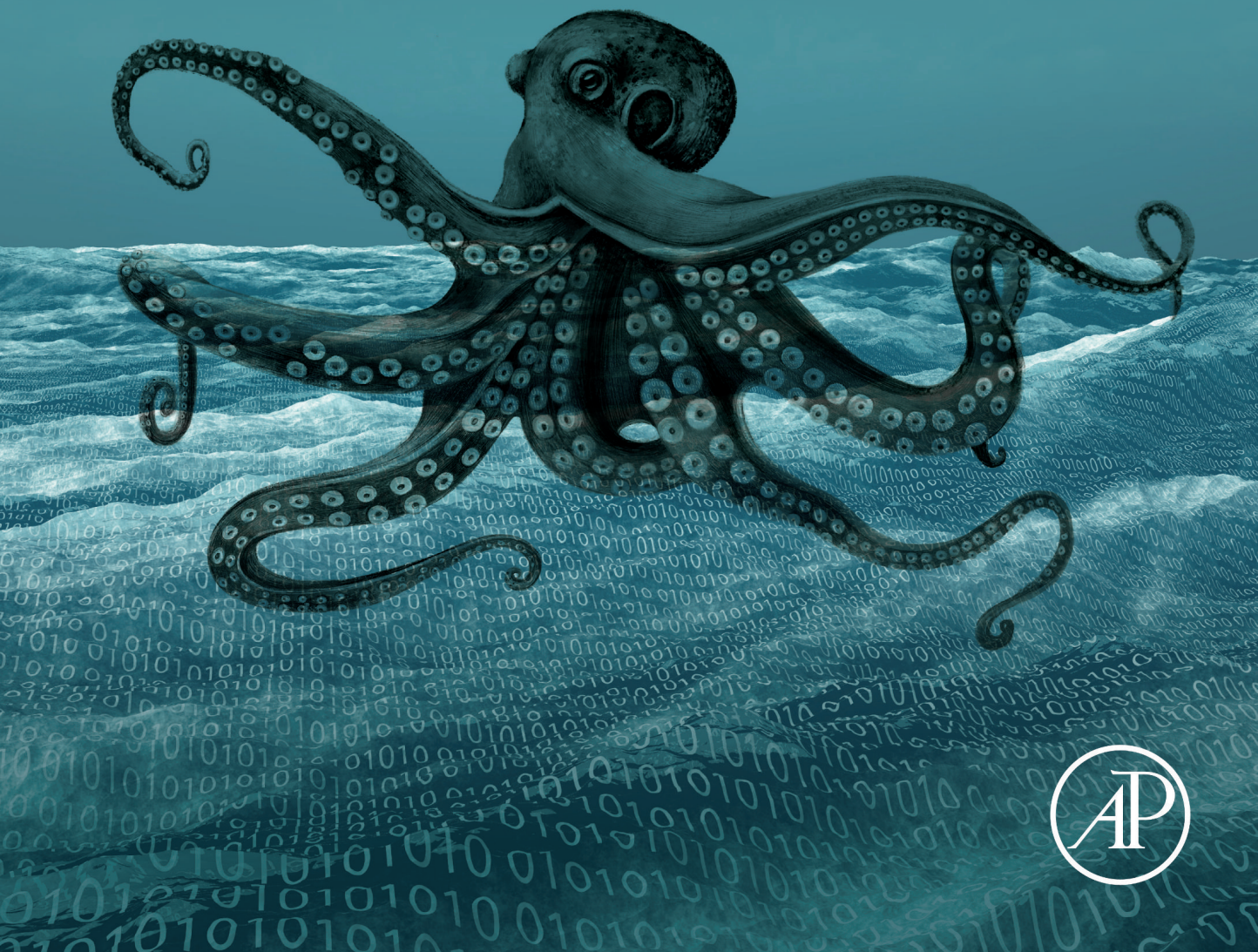
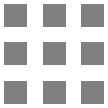
AI Assurance

Towards Trustworthy, Explainable, Safe, and Ethical AI

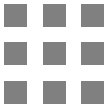
Feras A. Batarseh
Laura J. Freeman





AI Assurance

This page intentionally left blank



AI Assurance

Towards Trustworthy, Explainable, Safe, and Ethical AI

Edited by

Feras A. Batarseh
Department of Biological Systems Engineering (BSE)
College of Engineering (COE) & College of Agriculture and Life Sciences (CALS)
Virginia Tech
Blacksburg, VA, United States
Commonwealth Cyber Initiative (CCI) & National Security Institute (NSI)
Virginia Tech
Arlington, VA, United States

Laura J. Freeman
Department of Statistics
National Security Institute
Virginia Tech
Arlington, VA, United States

ACADEMIC PRESS
An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-323-91919-7

Publisher: Mara E. Conner
Acquisitions Editor: Chris Katsaropoulos
Editorial Project Manager: Emily Thomson
Production Project Manager: Fizza Fathima
Cover Designer: Mark Rogers

Typeset by VTeX

To our families

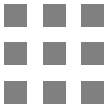
. . . and to all of you out there building safe and ethical AI systems.

Feras A. Batarseh and Laura J. Freeman

This page intentionally left blank

Contents

This page intentionally left blank



Contributors

Authors affiliations

Authors of this book come from the following institutions:

Commonwealth of Virginia Universities:

- Virginia Tech
- George Mason University
- University of Virginia
- University of Mary Washington
- Radford University

Other US Universities:

- University of California, Berkeley
- Georgetown University
- Missouri University of Science and Technology
- U.S. Air Force Academy
- Kansas State University
- Purdue University

International (Non-US) Universities:

- King's College, London, UK
- Aalto University, Espoo, Finland
- National Institute for Research in Computer Science and Automation (INRIA), Paris, France
- Military School of Science and Technology, Dhaka, Bangladesh

US Government Agencies:

- Federal Reserve Board
- Department of Health and Human Services

Industry:

- United Services Automobile Association
- Infinite Sum Modeling LLC
- Deloitte
- Sentara Healthcare

Contributors

Farid Ahmed

Nuclear Science and Engineering, Military School of Science and Technology, Dhaka, Bangladesh

Ayodeji Alajo

Nuclear Engineering and Radiation Science, Missouri University of Science and Technology, Rolla, MO, United States

Syed Bahauddin Alam

Nuclear Engineering and Radiation Science, Missouri University of Science and Technology, Rolla, MO, United States

Amira Al-Khulaidy Stine

George Mason University, Computational and Data Sciences Department, Fairfax, VA, United States

William Ampeh

Research and Statistics Division of the Federal Reserve Board, Washington D.C., United States

George Mason University, Fairfax, VA, United States

Ashita Anuga

Commonwealth Cyber Initiative, Virginia Polytechnic Institute and State University, Arlington, VA, United States

Feras A. Batarseh
Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States

Sindhu Bharathi
Infinite Sum Modelling LLC, Seattle, WA, United States

Sumathi Chakravarthy
Infinite Sum Modelling LLC, Seattle, WA, United States

Jaganmohan Chandrasekaran
Commonwealth Cyber Initiative, Virginia Tech, Arlington, VA, United States

Laura J. Freeman
Department of Statistics, National Security Institute, Virginia Tech, Arlington, VA, United States

Jay Gendron
Model the Cause, Chesapeake, VA, United States

Nitish Gorentala
Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, United States

Sai Gurrapu
Apple Inc., Cupertino, CA, United States

Flora Haberkorn
International Finance Division of the Federal Reserve Board, Washington D.C., United States

Hamdi Kavak
George Mason University, Computational and Data Sciences Department, Fairfax, VA, United States

Uma Krishnaswamy
International Finance Division of the Federal Reserve Board, Washington D.C., United States
University of California, Berkeley, Berkeley, CA, United States

Erik W. Kuiler
George Mason University, Fairfax, VA, United States

Dinesh Kumar
Institut National de Recherche en Informatique et en Automatique, Palaiseau, France
Nuclear Engineering and Radiation Science, Missouri University of Science and Technology, Rolla, MO, United States

William Franz Lamberti
Center for Public Health Genomics, University of Virginia, Charlottesville, VA, United States

Antoni Lorente
Department of Digital Humanities, King's College London, London, United Kingdom

Ralitsa Maduro
Sentara Healthcare, Virginia Beach, VA, United States
Virginia Wesleyan University, Virginia Beach, VA, United States

Connie L. McNeely
George Mason University, Fairfax, VA, United States

Anderson Monken
International Finance Division of the Federal Reserve Board, Washington D.C., United States
Georgetown University, Washington D.C., United States

Badri Narayanan
Infinite Sum Modelling LLC, Seattle, WA, United States

Shounkie Nawani
Infinite Sum Modelling LLC, Seattle, WA, United States

Minh Nguyen
Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

Kim Niewolny
Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

Ayorinde Ogunyiola
Purdue University, West Lafayette, IN, United States

Dominick J. Perini
Northrop Grumman, Denver, CO, United States

Brianna B. Posadas
Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

Abdul Rahman
Deloitte & Touche, LLP, Baltimore, MD, United States

Md Nazmul Kabir Sikder
Bradley Department of Electrical and Computer Engineering (ECE), Virginia Tech, Arlington, VA, United States

Andrei Svetovidov
Commonwealth Cyber Initiative, Virginia Tech, Arlington, VA, United States

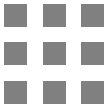
Amanda Tolman
Radford University, Radford, VA, United States

Hong-Linh Truong
Department of Computer Science, Aalto University, Espoo, Finland

Shoaib Usman
Nuclear Engineering and Radiation Science, Missouri University of Science and Technology, Rolla, MO, United States

Pei Wang
Microsoft Corporation, Redmond, WA, United States

Madison J. Williams
University of Mary Washington, Fredericksburg, VA, United States



A note by the editors

A group of AI experts congregated to write this book about assurance. Authors present assurance foundations (in Part 1), introduce new assurance formal methods (in Part 2), and provide examples of assurance applications in multiple domains (Part 3). Besides highlighting the importance and serious need for extensive evaluation of AI systems, AI Assurance provides a process towards ethical, explainable, fair, safe, secure, and trustworthy AI.

How to read this book?

If you are an AI researcher...

The goal of this book is to provide you with a foundation to understanding the conceptual, statistical, and theoretical challenges of AI assurance. We provide three literature review studies for bias and fairness, explainable AI, and outlier detection in Chapters 3, 4, and 7 to establish the state-of-the-science in AI assurance-related dimensions. Part 2 (methods) provides novel approaches to the assurance of AI systems of all kinds, including using causation, coordination, inference, and data management methods.

If you are an AI practitioner...

By reading this book, you will explore empirical studies on assumptions that influence algorithmic accountability and how they play out in practice (Part 3). Throughout the book, authors highlight the various factors that AI engineers negotiate when implementing AI; mostly in the following domains: economics (Chapter 11), healthcare (Chapter 12), engineering (Chapter 13), agriculture (Chapters 14 and 15), and technology policy (Chapter 16). AI assurance best practices are presented in Parts 1 (theoretical) and 2 (statistical).

If you are a policy maker...

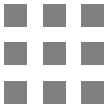
This book helps you identify potential methods for evaluating the use of AI algorithms at government in a liable manner. Methods related to AI for

public policy provide measures that can increase trust in AI systems and mitigate potential algorithmic harms through assurance. Begin with Part 3, and observe examples of evidence-based policy making. Additionally, take a look at the forewords by each section; they provide testimonies by executives trying to deploy AI in the public sector.

Lastly, if you are a non-technical AI enthusiast, we recommend that you begin by reading chapters: 1, 2, 5, and 6 before digging deeper into the inner workings of AI methods presented in other chapters.

Feras A. Batarseh
Associate Professor, Department of Biological Systems Engineering (BSE)
& Commonwealth Cyber Initiative (CCI)
Affiliate faculty, Center for Advanced Innovation in Agriculture (CAIA)
& National Security Institute (NSI)
Virginia Tech, Arlington, VA, United States

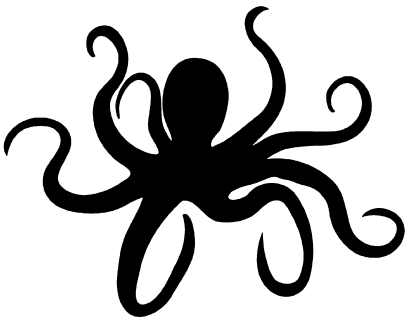
Laura J. Freeman
Research Associate Professor, Department of Statistics & National Security
Institute (NSI), Virginia Tech, Arlington, VA, United States



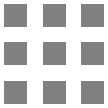
A note on the book cover

AI assurance is an octopus in a sea of data; it is required to be intelligent, adaptive, and accessible to all parts of its ecosystem. Octopuses are highly agile and intelligent carnivores; they can store long- and short-term memory information, can quickly learn from shapes and patterns of sea objects, have been reported to practice observational learning, and are known for building shelters for protective measures against adversaries.

Reference: Nuwer, R., "An Octopus Could Be the Next Model Organism," Scientific American, March 2021.



This page intentionally left blank



Foreword 1

I currently have the pleasure of leading the Commonwealth Cyber Initiative (CCI), a major investment by the Commonwealth of Virginia in research, workforce development, and innovation. We are more than 300 researchers from across Virginia with a focus on the intersection between cybersecurity, autonomy, and intelligence. The evolving field of Artificial Intelligence (AI) Assurance is at the very center of this intersection.

AI is in many of the products and services that we use on a daily basis, often without our being completely aware of it. There is AI in our smartphones, making decisions about how we connect to various networks and technologies, and in our social media, deciding what ads and posts we are exposed to. As cyber-physical systems, such as drones or robots, become more prevalent, they will be largely driven by AI.

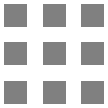
Some of these uses of AI affect critical services provided by a company, or involve health and safety issues (for example, in the use of AI in autonomous or assisted driving vehicles, another active area of CCI research and innovation). For these applications, treating AI as a black box—even one that produces great results in the vast majority of cases—is not enough. There must be assurances that the system is trustworthy, unbiased, explainable, ethical, and fair.

Other systems, such as next-generation communication networks, have traditionally relied on heuristics and experienced engineers, but are becoming too complex to be managed by humans. The need for AI is clear, but explainability and/or interpretability are key for widespread adoption.

This book gathers an impressive multi-disciplinary group of authors who introduce the fundamental goals and principles of AI Assurance and review the state-of-the art in the area, describe some of the main techniques adopted for assurance, and outline an array of applications in diverse areas, from healthcare to precision agriculture.

Ours is fast becoming a world of AI embedded into virtually everything. The potential for an increase in productivity and quality of life is great, as long as it is done responsibly. AI assurance will be even more critical in this new world.

Luiz DaSilva
Bradley Professor of Cybersecurity, Virginia Tech
Executive Director, Commonwealth Cyber Initiative



Foreword 2

Artificial Intelligence (AI) is here to stay and will continue to affect virtually every industry and every human being in the developed world. The healthcare sector, with its abundance of data, will be particularly affected by the advancement of AI and the big data paradigm shift. Together, they will drive innovation in healthcare, to include advancement in biomedical research, prevention of diseases, testing of life-saving drugs and vaccines, as well as spearheading the development of innovative medical devices. AI will also increase productivity exponentially, empowering healthcare providers to increase the volume, efficiency, and quality of delivery.

There are numerous use cases for AI to have a major impact on healthcare. For instance, it can augment human tasks and abilities by aiding in clinical decisions, supporting judgment and increasing treatment efficiency. Health professionals can access an abundance of data around diagnostic resources and research with new velocity. In the field of clinical research, AI can aid researchers and scientists by providing them with the ability to solve complex, global health challenges with the right data that is difficult to uncover with human analysis alone. AI algorithms can sort through large number of datasets unimaginable to human's ability with a high degree of accuracy and, in some cases, without bias. Health organizations are already exploring AI-based projects to discover health solutions across research and medical settings. The second area is in targeted diagnostics. Diagnosis and treatment of disease has been a core function of AI in healthcare. A patient can present symptoms that may require precise detection, diagnosis, treatment plan and an outcome prediction. The ability for AI to learn from the data provides the opportunity for improved accuracy based on feedback and enforcement responses. There are many research studies that have shown that AI can perform as well as or better than humans at healthcare tasks. The implementation of AI can also provide early detection by pinpointing risk alerts. This extends to telehealth tools that

can diagnose patients at the edge and into homes to help treat and prevent medical situations, while reducing hospital visits and readmissions. The third area is drug testing and drug discovery. Drug testing can be a lengthy process to make sure the proposed drugs will not pose harm to the consumers. AI can predict the attributes of toxicity of new compounds using data from past tests and experiments. These processes are extremely costly and time consuming; leveraging AI can help expedite the process with better precision.

AI gives us a tremendous amount of power to shape our healthcare of the future but it comes with a tremendous amount of responsibility, challenges, and concerns. There are also varieties of ethical implications around the use of AI in healthcare. If we allow AI to make or assist with health decisions, it raises issues of transparency, accountability, and privacy. Mistakes are to be expected by AI systems in patient diagnosis and treatment, and it may be difficult to establish accountability for them. AI in healthcare may also be subject to algorithmic bias, if the datasets are not inclusive of all races, genders, backgrounds, ages, and ethnicities.

It is imperative that healthcare institutions, government, and regulatory entities establish structures to continuously create and enforce rules to limit negative implications. Additionally, we must ensure that we are doing our best to bridge the digital divide and not leave people behind. Datasets are mostly being captured by people in developed countries that have access to technology. Coincidentally, characteristics and norms of people in countries with lack of access to the internet will not have footprint online, and thus not represented in the analysis. The result is that AI-generated information will have unintentional bias.

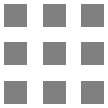
For health organizations to embark on the AI journey, they must apply some guardrails. By designing AI solutions with these principles in mind, they can protect against unintended consequences and promote ethical AI decision-making. They must consider privacy, legal, ethics, security, trustworthiness, and free from bias.

The deployment of AI in the health sector holds the promise to improve efficiency, effectiveness, safety, fairness, welfare, transparency, and other economic and social goals. Health organizations have shown efforts to date

to demonstrate its desire and commitment to fully realize the benefits of AI. AI provides capabilities that solve complex mission challenges and generate AI-enabled insights to inform efficient programmatic and business decisions, while removing barriers to innovation. We have an opportunity to shape our next generation story. We can create a new story on how we want to live, socialize, work, and play. We need to take responsibility at every level of society to adapt to these technological challenges and changes, with the intent of redefining what is like to be a human. Let us look at AI as a force for good if implemented correctly. We have the opportunity to write our own autobiography. Let us help meet the basic needs of all humans and write our future.

Oki Mek
Chief Artificial Intelligence Officer (CAIO)
U.S. Department of Health and Human Services (HHS)

This page intentionally left blank



Foreword 3

Artificial Intelligence (AI) is recognized as one of the most powerful technologies in generations with the potential for transformation impacts to America's prosperity, welfare, and national security. In its final report published in March 2021, the National Security Commission on Artificial Intelligence (NSCAI) calls on the U.S. government to "expand our conception of national security and find innovative AI-enabled solutions" to harness its power for our nation's defense. In fact, this is already happening. AI has become ubiquitous: machine and deep learning are now standard techniques in data analysis. Organizations in industry and government alike have launched enterprise data management and software development strategies to accelerate AI innovation and adoption. Within the United States' nuclear security enterprise, advances in computing, intelligent methods, and new data sources present new opportunity to enhance capability to detect, monitor, and verify global activities to develop and proliferate nuclear weapons.

Working with allies and partners around the globe, the United States employs nuclear nonproliferation and arms control to reduce the dangers posed by nuclear weapons. The White House's *Interim National Security Strategic Guidance* published in March 2021 recognizes the urgency and key role of nuclear nonproliferation and arms control in the United States' strategic stability. Central to nuclear nonproliferation and arms control are technologies and science-based methods to detect activities by state and non-state actors to develop or acquire nuclear weapons-usable capabilities and assess nascent or extant nuclear weapons programs. The United States seeks to leverage AI to expand nuclear proliferation detection and enable the detection of nuclear threats earlier than ever before, which affords more options for intervention.

The missions and decisions enabled and informed by nuclear proliferation detection technologies are high-consequence, highly technical, and

executed under challenging operational conditions. Designing intelligent systems that are useful for nuclear proliferation detection requires the use of clever assurance and validation techniques beyond the standard methods typically used in AI and machine learning. Nuclear nonproliferation decision-makers, operators, and national security analysts require transparency in the processes and workflows used to manipulate and analyze data to ensure they can trust and act on the results. The pressing nature of these missions demands that decisions be made even under uncertain conditions: intelligent systems must perform predictably when exposed to out-of-distribution data or when the availability of a data source changes. In addition to conventional definitions of trustworthy and ethical AI, intelligent systems used in nuclear proliferation detection are accountable to laws of science and engineering-defined constraints, such as physical and chemical properties. This presents both a requirement for validation and an opportunity to constrain and inform learning.

Thus, this text presents a myriad of techniques aimed at evaluating the assurance of AI systems. The material of this book is essential in realizing the charge of the *NSCAI Final Report*: to realize an "AI-ready national security enterprise by 2025". AI assurance is only achievable using techniques with strong scientific and rigorous mathematical underpinnings that are selected based on the challenges of the mission for which the intelligent system will be used.

Leveraging the clever techniques presented in this text, it is possible to build AI-enabled technologies that outperform many existing capabilities and that can be adopted and integrated into operational national security mission to make the world safer.

Angela M. Sheffield
U.S. Air Force Academy and Kansas State University

PART 1

Foundations of AI assurance

This page intentionally left blank

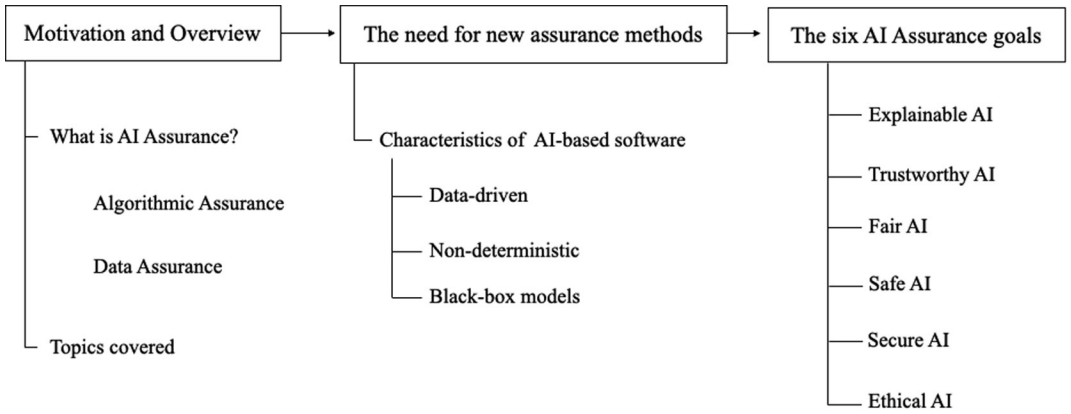
1

An introduction to AI assurance

Feras A. Batarseh[a], Jaganmohan Chandrasekaran[b], and Laura J. Freeman[c]

[a]*Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States* [b]*Commonwealth Cyber Initiative, Virginia Tech, Arlington, VA, United States* [c]*Department of Statistics, National Security Institute, Virginia Tech, Arlington, VA, United States*

Graphical abstract



Abstract

In this chapter, we present a brief introduction about the concept, dimensions, and challenges of AI assurance. The goal is to lay the path for the concepts presented in this book. AI is often assessed by its ability to consistently deliver accurate predictions of behavior in a system. A critical, often overlooked, aspect of developing AI algorithms is that performance is a function of the task the algorithm is assigned, the domain over which the algorithm is intended to operate, and changes to these elements in time. These parameters and their constituent parts form the basis which makes assuring AI a challenge. Algorithms need to be characterized by understanding the factors that contribute to stable performance across an operational environment (e.g., no dramatic perturbation by

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00013-5

small changes and/or no effects measurable over time). This chapter presents a high-level introduction to AI Assurance and points readers to related areas of interest in the book.

Keywords

AI assurance, testing & evaluation, responsible AI

Highlights

- An introduction to AI assurance
- Topics covered and questions answered in this book
- Book's main thrusts and high-level summaries
- The need for AI assurance

1.1 Motivation and overview

To accurately and consistently predict behaviors in systems, AI systems require data for training and testing the outcomes. The iterative process of improving accuracy and precision in developed models involves trade-offs in performance, data quality, and other environmental factors. AI's predictive power can be impacted through changes in the training/testing data, the model, and its context. In this chapter, we discuss sources of change captured within the operational context (Brézillon and Gonzalez, 2014) of an AI's execution and that is often attributed to inconsistencies in AI systems. Model and data changes are discussed in this book, especially those related to concept drift in applications with examples of how these inconsistencies emerge (Žliobaitė et al., 2016; McPherson, 2021; Tucker, 2021).

The need for assurance was reinforced in a recent report by the US National Security Commission on Artificial Intelligence (NSCAI), which proposed that government agencies, such as the National Institute of Standards and Technology (NIST) and the National Institute of Food and Agriculture (NIFA) "should provide and regularly refresh a set of standards, performance metrics, and tools for qualified confidence in AI models, data, training environments, and predicted outcomes" (NSCAI, 2021). The government, industry, and academia are required to collectively advance the

AI community in establishing these resources and standards. Accordingly, this book's motivation is to provide a vision for the path forward and serve as a foundational and theoretical manifesto to the assurance of AI systems.

1.1.1 Book content

The incremental testing movement was an afterthought in the software engineering world. If we aim to learn from that experience, however, we ought to develop assurance incrementally and as part of the AI lifecycle. Therefore assurance should not be treated as a separate component while developing AI systems; instead, it should be a part of the incremental learning process of any agent, environment, or algorithm. **Chapter 2** of this book discusses the notion that "the process of ensuring fair, unbiased, and ethical AI needs to be a continuous endeavor, making of AI assurance a process and not a goal;" it also includes a valuable exploration of the areas of generalization and other major assurance challenges, such as the control problem, value loading, and human-AI alignment.

A well articulated process and a clearly defined set of metrics to categorize and measure the maturity of assurance could go a long way in establishing a common understanding of these systems' dependability. Similar process structures, e.g., the capability maturity model integrated (CMMI), have been employed to measure an organization's ability to produce high quality software systems. The advantage of such a model is that it encourages all stakeholders to agree on a set of metrics and processes to measure the quality of the AI systems being produced and deployed. It also shows the path to achieve a gradually higher level of assurance following a consensus set of criteria. We believe that a similar set of metrics and process under a maturity model framework will not only streamline AI systems' development efforts, it will also foster sharing of implementation experiences and best practices. However, such standards should be defined based on AI-related metrics, for instance, **Chapter 3** presents a rich overview of statistical methods and foundational metrics to measuring assurance of AI systems, with focus on explainability and interpretability. However, if we consider assurance goals, such as explainability, fairness, and trustworthi-

ness, trade-off decisions have to be made. Algorithms that are more complex (such as neural networks) tend to be less interpretable and prone to different kinds of bias for instance. As reported by Gunning and Aha (2019), the performance of AI algorithms is inversely proportional to the explainability of the model's decision. Accordingly, **Chapter 4** presents bias reduction methods and compares them in terms of the overall validation of AI systems.

As AI is getting adopted across all domains, assuring AI systems is becoming a matter of national security; it has effects on manufacturing, cyber-physical systems, the economy, healthcare, government, and many other sectors; **Chapter 5** introduces potential short- and long-term global impacts of AI assurance. Accomplishing assurance is a complicated endeavor nonetheless; **Chapters 6**, **8**, and **10** present detailed frameworks and lifecycles that can be used to establish a process for assurance within any domain, by applying *algorithm assurance* concepts, such as inference, causality, resilience, and elasticity. *Data assurance,* however, is another critical dimension in AI assurance (Kulkarni et al., 2020); **Chapters 7** and **9** provide answers and recommendations on data wrangling methods for improved model outcomes, as well as addressing outlier detection issues in training data and their effects on the outcomes of learning algorithms. The last part of the book presents a variety of applications and illustrates the need for assurance in many sectors such as Economics (**Chapter 11**), Healthcare (**Chapter 12**), Engineering (**Chapter 13**), Agriculture (**Chapters 14** and **15**), and Public Policy (**Chapter 16**).

1.2 The need for new assurance methods

Recent advancements in AI have demonstrated the potential of AI-based software systems in successfully performing tasks that generally require human-level intelligence. A survey by Batarseh et al. (2021) recommends a set of assurance goals, provides a new comprehensive definition for AI assurance, and suggests that AI-based software systems are rapidly adopted across various domains. An AI-based software system consists of one or more machine learning models that are used to perform intelligent tasks, such as object identification, pedestrian detection, speech translation, and

decision support. In the AI engineering lifecycle, developing a model is a multi-step process. One of the critical initial steps is algorithm selection. An AI framework, such as sci-kit learn, Tensorflow, or Pytorch, consists of a collection of off-the-shelf AI algorithms. The AI algorithm analyzes the dataset, infers, and learns the hidden patterns, and derives a decision logic on receiving the input. This activity is referred to as the training phase, and the derived decision logic is referred to as a trained AI model (Chandrasekaran, 2021; Felderer and Ramler, 2021). In the training phase, multiple assurance challenges could be faced, such as data bias, data incompleteness, dark data, or data collection inconsistencies. Despite the promising potential demonstrated by AI-based software systems, they are error-prone and tend to fail once deployed in real-world environments (Lee, 2016; Dastin, 2018; Vincent, 2020; Mitchell, 2021; E.Boudetter, 2021). Such failures can have serious consequences, including fatal consequences in safety-critical domains (Cellan-Jones, 2020; Newman, 2021). However, assurance, testing, validation, and verification of systems is not a new problem, the software engineering community has made major progress and multiple conclusions on these fronts, some of which could be very useful for AI, whilst others are not related at all. In the remaining of this section, we argue against recycling existing assurance methods, and present the case for a new set of AI assurance methods. In the software development lifecycle, testing is performed before the software system is released. The objective of the testing activity is to ensure that the software system will behave as intended. According to ISO/IEC/IEEE 29119-1:2013 (IEEE_Software_Testing, 2013), the primary goals of software testing are to provide information about the quality of the test item and any residual risk in relation to how much the test item has been tested, to find defects in the test item prior to its release for use, and to mitigate the risks to the stakeholders of poor product quality. Poor software quality can have an adverse effect and cause severe damages to its stakeholders. A report in synopsis states that, in 2021, the software glitches in the US cost around an estimated $2 trillion (Armerding, 2021). Testing is a complex yet essential activity in the software development lifecycle. Over the years, several approaches and methodologies have been developed to effectively test and release software systems. How-

ever, they are tailored towards testing and evaluation of traditional software systems, and not AI. In traditional software systems, the decision logic is written by humans based on the requirements provided by stakeholders. More importantly, decision logic is deterministic, that is, for a given input, the system is guaranteed to produce the exact output at each execution. In contrast, AI systems derive their logic from a training dataset, and in most cases, the algorithms in an AI-based software system behaves in a stochastic manner. Furthermore, an AI software system shall exhibit a change in its behavior with different data, different contexts, and different users; all of which obviously exacerbate the assurance challenge (Freeman, 2020). Therefore the behavior of an AI-based software system is influenced by a combination of factors, all requiring assurance. Additionally, in the case of traditional software systems, the decision logic is derived based on the requirements. Hence, the test cases are generated based on the business requirements, and each test case shall have a predetermined/predefined set of outputs. On the contrary, in AI-based software systems, there are no written requirements in the traditional sense. Instead, the AI model derives the logic from the training dataset (through supervised or unsupervised training processes). Therefore AI-based software systems suffer from the test oracle problem (Weyuker, 1982; Murphy et al., 2007). That is, in most cases, the intended system behavior for a test case can hardly be predefined. In other words, the exact intended behavior of an AI software system is not fully known until the scenario occurs in real-time (such as in reinforcement learning scenarios).

An AI-based software system with a higher prediction accuracy (closer to a 100% accuracy) is expected to guarantee an error-free behavior, also, they are expected to make objective, impartial decisions. On the contrary, in most cases, when deployed in real-world conditions, AI-enabled software systems can inadvertently result in discriminatory behavior. For example, an AI algorithm used by a major US-based healthcare institution to identify patients for extra care through an intensive care management program appeared to be discriminatory against patients of African-American ancestry (Strickland, 2019). The root cause for such unintentional yet discriminatory behavior could be attributed to the inherent bias in the dataset used to train

the AI model used in the AI-based software system. Issues reported in Buolamwini and Gebru (2018); Strickland (2019); Caliskan (2021); Zang (2021) indicate that evaluating the quality of AI-based systems requires determining beyond the correctness of the AI systems. As AI-based software systems are data-intensive, in addition to the correctness, it is essential to test for bias and variance in the system (to identify over or under-fitting issues). From an assurance standpoint, it is imperative to develop standardized assurance methods that are capable of detecting and mitigating any biased behavior before AI-based software systems are deployed.

Furthermore, in the case of a traditional software system, on executing a test case, a deviation of the *observed* behavior from the *expected* behavior is considered as failure. However, in the case of an AI-based software system, the correctness of system behavior is evaluated based on the prediction accuracy (a statistical score) of the AI system (Zhang et al., 2020; Riccio et al., 2020). A statistical score (for example, correct predictions/total predictions) is calculated over a test dataset. A model achieving a higher accuracy is considered one of higher quality. Also, the acceptable threshold of the prediction accuracy score varies across domains, users, and models. It follows that, as this book presents, assurance of AI systems could be domain-specific or domain-dependent, model-specific or model-agnostic, but is certainly needed in all cases, scenarios, and deployments. For traditional software systems, in most cases, the root cause for an unexpected behavior (failure) can be localized to a segment in the source code. However, when an AI system exhibits a failure, it can be caused by either by the training dataset, missing data, outliers, choice of hyper-parameters, or by the trained model and its architecture (Batarseh and Gonzalez, 2018). For example, as reported in Wiggers (2021), the inherent bias in the training dataset results in a discriminatory AI software system. In other cases, even the choice of an AI algorithm can be attributed to unexpected behavior (Yee et al., 2021). Given the fundamental differences between a traditional software system and an AI-based software system, and the quality assurance challenges that arise from these differences, it is vital to develop assurance methods that are especially tailored and best suited to assess and evaluate AI-based systems, a notion that is covered in this book.

1.3 Conclusion

As reflected by this book, AI assurance is based on a set of trade-offs. An AI model that exhibits better performance is generally a black-box, and their reasoning (or) decision-making process is not easily understandable to the users. From an assurance standpoint, in addition to evaluating a model's correctness (accuracy), it is essential to understand why a model makes a specific decision. Despite the prediction capabilities, as the reasoning behind a model's decision is largely opaque, it leads to a lack of trustworthiness among the users. From a quality assurance perspective, it is essential to develop approaches and tools that will generate fair outcomes that are secure, safe, and easily understandable to all stakeholders (AI engineers, end-users, business owners, data scientists) involved in the process. Accordingly, we provide an updated definition of AI assurance, which is an extension to the one presented in Batarseh et al. (2021); this definition is adopted across this book: AI assurance is a process that is applied at all stages of the AI engineering lifecycle, ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy, and explainable to a layman, resilient against adversaries, robust within its domain, ethical in the context of its deployment, unbiased in its learning, and fair to its users.

References

Armerding, T., 2021. What is the cost of poor software quality in the US? https://www.synopsys.com/blogs/software-security/poor-software-quality-costs-us/.

Batarseh, F.A., Freeman, L., Huang, C.H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8, 1–30.

Batarseh, F.A., Gonzalez, A.J., 2018. Predicting failures in agile software development through data analytics. Software Quality Journal 26, 49–66.

Brézillon, P., Gonzalez, A.J., 2014. Context in Computing: a Cross-Disciplinary Approach for Modeling the Real World. Springer.

Buolamwini, J., Gebru, T., 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. PMLR, pp. 77–91.

Caliskan, A., 2021. Detecting and mitigating bias in natural language processing. https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/.

Cellan-Jones, R., 2020. Uber's self-driving operator charged over fatal crash. https://www.bbc.com/news/technology-54175359.

Chandrasekaran, J., 2021. Testing artificial intelligence-based software systems. Ph.D. thesis.
Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
E.Boudetter, N., 2021. Tesla says autopilot makes its cars safer. crash victims say it kills. https://www.nytimes.com/2021/07/05/business/tesla-autopilot-lawsuits-safety.html.
Felderer, M., Ramler, R., 2021. Quality assurance for AI-based systems: overview and challenges (introduction to interactive session). In: International Conference on Software Quality. Springer, pp. 33–42.
Freeman, L., 2020. Test and evaluation for artificial intelligence. Insight 23, 27–30.
Gunning, D., Aha, D., 2019. Darpa's explainable artificial intelligence (xai) program. AI Magazine 40, 44–58.
IEEE_Software_Testing, 2013. Iso/iec/ieee international standard - software and systems engineering –software testing –part 1:concepts and definitions. In: ISO/IEC/IEEE 29119-1:2013(E), pp. 1–64.
Kulkarni, A., Chong, D., Batarseh, F.A., 2020. Foundations of data imbalance and solutions for a data democracy. In: Data Democracy. Elsevier, pp. 83–106.
Lee, P., 2016. Learning from Tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.
McPherson, S., 2021. Fixing 'concept drift': retraining AI systems to deliver accurate insights at the edge. https://gcn.com/articles/2021/07/16/ai-concept-drift.aspx.
Mitchell, R., 2021. Tesla's handling of full self-driving bug raises alarms. https://www.latimes.com/business/story/2021-11-03/teslas-handling-braking-bug-in-public-self-driving-test.
Murphy, C., Kaiser, G.E., Arias, M., 2007. An approach to software testing of machine learning applications.
Newman, R., 2021. It's time to notice Tesla's autopilot death toll. https://news.yahoo.com/its-time-to-notice-teslas-autopilot-death-toll-195849408.html.
NSCAI, 2021. Chapter 7 - NSCAI final report. https://reports.nscai.gov/final-report/chapter-7/.
Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., Tonella, P., 2020. Testing machine learning based systems: a systematic mapping. Empirical Software Engineering 25, 5193–5254.
Strickland, E., 2019. Racial bias found in algorithms that determine health care for millions of patients. https://spectrum.ieee.org/racial-bias-found-in-algorithms-that-determine-health-care-for-millions-of-patients/.
Tucker, B., 2021. Managing the risks of adopting AI engineering. https://insights.sei.cmu.edu/blog/managing-the-risks-of-adopting-ai-engineering/.
Vincent, J., 2020. AI camera operator repeatedly confuses bald head for soccer ball during live stream. https://www.theverge.com/tldr/2020/11/3/21547392/ai-camera-operator-football-bald-head-soccer-mistakes.
Weyuker, E.J., 1982. On testing non-testable programs. Computer Journal 25, 465–470.
Wiggers, K., 2021. Employees attribute AI project failure to poor data quality. https://venturebeat.com/2021/03/24/employees-attribute-ai-project-failure-to-poor-data-quality/.

Yee, K., Tantipongpipat, U., Mishra, S., 2021. Image cropping on Twitter: fairness metrics, their limitations, and the importance of representation, design, and agency. arXiv preprint arXiv:2105.08667.

Zang, J., 2021. Solving the problem of racially discriminatory advertising on Facebook. https://www.brookings.edu/research/solving-the-problem-of-racially-discriminatory-advertising-on-facebook/.

Zhang, J.M., Harman, M., Ma, L., Liu, Y., 2020. Machine learning testing: survey, landscapes and horizons. IEEE Transactions on Software Engineering.

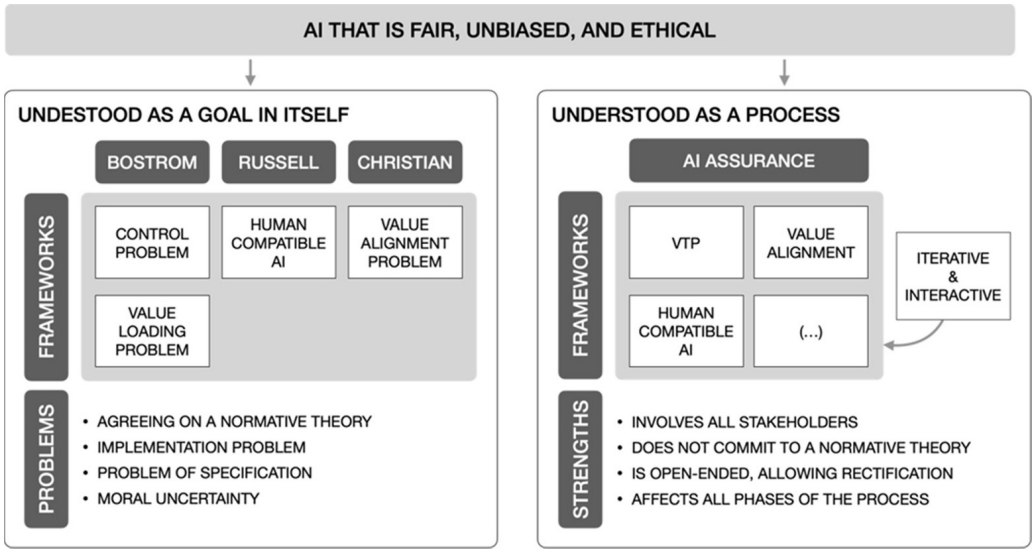
Žliobaitė, I., Pechenizkiy, M., Gama, J., 2016. An overview of concept drift applications. In: Big Data Analysis: New Algorithms for a New Society, pp. 91–114.

2

Setting the goals for ethical, unbiased, and fair AI

Antoni Lorente
Department of Digital Humanities, King's College London, London, United Kingdom

Graphical abstract



Abstract

The main goal of AI assurance is to ensure that AI systems are, among other things, ethical, unbiased, and fair. In this chapter, three different approaches to the value alignment problem, i.e., how to ensure that an AI's decisions and behaviors are aligned with our values, are introduced. The chapter claims that AI assurance provides a shared vernacular and a formal framework to meaningfully apply the strategies to deal with the value alignment problem above, motivating several questions that are fundamental to such alignment. A brief overview of three different normative theories pinpoints the dilemmatic nature

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00014-7

of defining "the good," justifying in turn the need to tackle the problem of implementation, specification, and moral uncertainty. It is argued that even though behavior-based learning allows deferring some of these questions, for AI assurance to attain its goals—both now and in the future—the process of ensuring fair, unbiased, and ethical AI needs to be a continuous endeavor, making AI assurance a process and not a goal.

Keywords

Value alignment problem, AI ethics, specification, implementation, uncertainty, CIRL, Artificial Intelligence

Highlights

- This chapter introduces AI assurance as a process that enables fair, unbiased, and ethical AI
- Ethical interpretations are explained via the problem of aligning AI systems with our values and interests
- The embrace of behavior-based value learning methods motivates the necessity to further explore AI assurance as a crucial actor within AI systems development

2.1 Introduction and background

AI assurance is a field of research entrusted with a crucial task: ensuring that the development and adoption of advanced AI systems does not jeopardize the fundamental pillars on which our society stands. Such assurance involves, consequently, discussions about development, transparency, commercialization, regulation, control, or use, which need to be addressed at multiple levels of abstraction: from the most fundamental mathematical formalisms to the complexity of everyday language. The aim of AI assurance is thus to make sure that any AI system being developed and deployed produces outcomes that are "valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users" (Batarseh et al., 2021).

The strength of the process suggested here is that it applies throughout the whole range of technologies that fall beneath the umbrella term "artificial intelligence". It is a claim that compels any and all AI systems to

satisfy a bare minimum of requirements for them to be acceptable not only in technical or commercial terms, but also social, ethical, and legal. AI assurance thus provides, on the one hand, the framework, while on the other, the vernacular to meaningfully assess each step of both the development and adoption process of AI systems for them to be fair, safe, unbiased, and ethical.

AI systems are increasingly embedded in our social context, and regardless of the technical brilliance that underpin them, the possible futures that AI opens up to us are both staggering and uncanny. Original and complex AI systems, capable of undertaking tasks that would otherwise be unfathomable, are many times obscure to the general public, justifying the grounds for anxiety. But once the emotion wears off, the difficult and often times profound philosophical questions remain. How should we align AI systems with our values? How similar to natural intelligence is artificial intelligence? While some of the questions are deeply metaphysical, especially those about conscience, free will, agency, and autonomy, many others pick on ethical dilemmas that have governed the progress of philosophy for the last millennia.

AI has evolved via different approaches to learning. However, one of the main problems when training a machine learning algorithm is **generalization,** or the capacity of a given model to adapt to new datasets. Intuitively, machine learning is the field of study that gives computers the ability to learn without being explicitly programmed to Samuel (1959). Such learning is materialized via a model about a given portion of the world, which is articulated by a hypothesis that is inferred from a dataset. If such dataset is previously partitioned into labeled categories, we call it **supervised learning**. Otherwise, the so-called learning is **unsupervised learning**. Last, if the agent conditions its action to the reward it receives, this is called **reinforcement learning** (RL).

The problem of generalization is troublesome for several reasons. First, it limits the capacity of an AI system to be used beyond the training dataset. But second and more importantly, it increases the chances of misinterpreting, or misclassifying elements from the real world. Given the growing impact of AI algorithms on our everyday lives, trying to align AI systems

with our values is a crucial task for researchers. It is in this sense that philosophical approaches to AI provide a meaningful background to reformulate technical problems as ethical and concrete philosophical problems.

This chapter begins with an overview of three main formulations of the value alignment problem in AI. Section 2.1.1 introduces Nick Bostrom's concerns regarding the control problem and the value-loading problem in the context of the existential risks that AI entails. Section 2.1.2 briefly discusses Stuart Russell's defense of human-compatible AI, paying special attention to his proposal of abandoning the standard model of AI, as well as to the possibilities that assistance games, such as cooperative inverse reinforcement learning bring. Section 2.1.3, discusses Brian Christian's value alignment problem, which provides meaningful insights regarding the role of training data and objective functions in developing aligned AI systems. In Section 2.1.4, AI assurance is interpreted as a process that provides a shared vernacular and a formal framework to implement the discussions above.

The second part of this chapter is rooted in some fundamental aspects regarding the development of safe and ethical AI. Section 2.2.1 introduces three of the most important normative theories in ethics: duty-based deontology, utilitarianism, and virtue ethics. After this, the implementation problem is discussed in Section 2.2.2, considering two possible strategies to develop a moral sense in a machine: a top-down and a bottom-up approach. Then, in Section 2.2.3, some problems related to the nature of intentional statements are introduced, focusing in particular on the problem of specification and the role of moral uncertainty.

The chapter concludes insisting on the idea that for AI assurance to succeed in its goals, the process of ensuring that AI systems are ethical, unbiased, fair, and safe needs to be an iterative, interactive, and deliberative process. Thus AI assurance reformulates AI ethics not as a goal, but as a new relationship with technology.

2.1.1 Value-loading

In the book "Superintelligence: Paths, Dangers, Strategies", Nick Bostrom (2014) presents two problems that are crucial for AI assurance. On the one hand, what Bostrom calls "**the control problem**" raises the question bear-

ing on which principles should buttress a framework to harness an artificial general intelligence that could overtake us. On the other hand, and before the provisional nature of control mechanisms, "**the value-loading problem**" allows us to formalize the puzzle of having intelligent systems whose values are aligned with ours. The two sections that follow provide a brief outline.

2.1.1.1 The control problem

Both the control and the value-loading problem are raised from the perspective of the **existential risk** that an AI—a general, superintelligent one—could entail in the longer run for humanity. And even though AI assurance and machine ethics are primarily concerned with current developments (i.e., those related to machines that are still far inferior to humans in terms of general intelligence), it is both intrinsically and instrumentally enriching to engage in the exercises that Bostrom proposes. The idea of an artificial general intelligence (or AGI) has motivated relevant interdisciplinary research agendas that have contributed both to prevent such outcome and to ensure better systems. Moreover, thinking about this existential risk not only allows us to consider the long-term consequences of our current decisions, it also puts into perspective our ultimate goals. It is, perhaps, because of this that Bostrom's work has been so influential.

But how could we control an AGI? To answer this question, Bostrom introduces the possibility of an artificial intelligence explosion (or a putative process in which a moderately intelligent agent improves radically until reaching a superhuman level of intelligence via recursive self-improvement) (Bostrom, 2014: 408). The main goal of the control problem is to achieve a "controlled detonation" of such intelligence (Bostrom, 2014, p. 155). To do so, the discussion targets which methods could be used to ensure that such agent realizes the sponsor's goals.

Bostrom proposes several methods to ensure a controlled detonation, which can be grouped into two categories: **capability control** and **motivation selection**. Capability control methods include the following:

- *Boxing*: This process consists of containing (either physically or informationally) the artificial intelligence.

- *Incentive methods*: These are methods that place the agent in an environment where it finds instrumental reasons to behave consistently with the designer and developer's interests.
- *Stunting*: Stunting consists of limiting the system's capabilities or access to information.
- *Tripwires*: These are occult diagnostic tests that shut down the AI if it identifies possible dangerous behaviors.

On the other hand, motivation selection methods include

- *Direct specification*: Direct specification is an attempt to explicitly define —either via rule-based or consequentialist principles—a set of rules or values to align the agent's behavior with our interests, domesticity, or placing the agent in a particular situation where specifying its behavior may be tractable.
- *Indirect normativity*: On the contrary, indirect normativity consists of specifying a process to derive norms.
- *Augmentation*: Augmentation consists of enhancing the capabilities of a system that already has an acceptable motivation system.

2.1.1.2 The value-loading problem

The measures proposed to control an AI explosion above, however, are only temporary. If we intend to have intelligent agents that are reliable and safe beyond highly limited settings, we must face what Bostrom calls the "value-loading problem". The difficulty of this problem lies in the impossibility to specify any and every possible situation the AI might encounter, and the type of behavior we would expect from it. Given this impossibility, motivation systems need to be specified more abstractly—in the form of a rule or a formula—that allows the agent to make a decision in new situations (Bostrom, 2014, p. 226). One possible way to do so could be via a utility function, which assigns a different value to each possible outcome according to a set of criteria, and then evaluates the best one. This function, and the behavior it would promote, should be aligned with our own goals. But this raises a grave problem: it is possible that such a utility function could only deal and achieve simple goals, or to select motivations in highly domesticated environments. It is in part because of this that explicit utility

functions seem to provide little help beyond highly simplified scenarios. In response to this, and to circumvent the hopeless endeavor of exhaustively specifying our values, Bostrom presents seven different value-loading techniques open for exploration.

- *Evolutionary selection*: Evolutionary selection captures the idea behind biological evolution. Simply put, evolution is a type of two-step search strategy that stochastically expands a given population with new candidates, which are then tested against an evaluation function and, ultimately, pruned. However the problem is that a search strategy could satisfy both conditions of evolutionary selection without satisfying our expectations in terms of value (Bostrom, 2014, p. 229).
- *Reinforcement learning*: A second option is to consider agents trained on reinforcement learning, where agents learn to maximize cumulative reward. In such cases, as the agent experiences, the evaluation function, or a function that tells the agent the value of its current state, is updated. This could be misunderstood as such function bears on a capacity to learn about values, but that is not the case: the agent is merely getting better at estimating instrumental values of reaching a particular state. Since the goal of the system is to maximize future reward, it seems unlikely that RL by itself can provide a solution to the problem at hand (Bostrom, 2014, pp. 230–1).
- *Associative value accretion*: This technique is grounded on the idea that humans are born with simple starting preferences and a set of dispositions to acquire new ones in response to experience. The presence of both is innate, but the development of the latter depends on one's life. The problem with this approach is that the mechanisms behind what representations are value-sensitive (e.g., someone's well-being) is not well understood. Therefore trying to mimic the human value-accretion process seems too difficult (Bostrom, 2014, pp. 231–3).
- *Motivational scaffolding*: Another possible approach is to build a primitive goal system into the seed AI with simple final goals explicitly coded. Once the AI develops more sophisticated representational capabilities, this initial scaffold is replaced by more complex goals that enable the system to bloom into a full superintelligence with a complex set of values.

However, since scaffold goals are not merely instrumental but final goals for it, the AI could oppose resistance to have them substituted. To deal with this danger, both control and motivation selection methods could be implemented (Bostrom, 2014, pp. 233–4).

- *Value learning*: Value learning rests on the idea that the system can use its intelligence to learn the values we want it to have. The AI would estimate an implicitly defined set of values based on an internal criterion—a scaffold, similarly to the motivational scaffolding technique—but its final goal would remain unchanged: what changes throughout the process is the system's belief about the goal. The key about this approach is that the system generates hypothesis about what values are worth pursuing, and changes the ponderation of each hypothesis according to evidence (Bostrom, 2014, pp. 235–7). Bostrom provides a possible formal approach to value learning. However, due to technical limitations and the need to implement a correct motivation before the superintelligence's explosion, this approach is described more as a research program than an available technique. However, cooperative inverse reinforcement learning, presented in the section below, will undoubtedly resonate with this proposal. (See Section 2.1.2.)
- *Emulation modulation*: Taking a different path towards developing intelligent agents, i.e., by means of brain emulation, opens up a different technique to tackle the value-loading problem. With emulations, the augmentation motivation selection method could be applicable. The idea rests on combining inherited goals, which are part of the emulation, with the equivalent for the system of psychoactive substances. However, and even though emulations seem easier to catalogue as moral agents than synthetic artificial intelligences, research on emulations can raise severe ethical concerns, such as the peril of augmenting an intelligence before testing and adjusting its final goals, or having it done by unscrupulous research groups (Bostrom, 2014, pp. 246–7).
- *Institution design*: The last technique that Bostrom considers is institution design. The core idea here is that whole brain emulations present the features necessary to be taken as a part of a composite system, which would in turn be subject to being constrained by institutions. Similar to

> what happens to companies, which are constituted by workers with their own agency, are many times considered to have their own autonomous agency. Similarly, by means of designing the appropriate institutions, AI systems embedded in this conglomerate would align their behavior consistently with the goals set by the governing institutions via whatever constraints they enacted (Bostrom, 2014, pp. 247–253).

All these techniques, including control and motivation selection methods, are the testimony of an attempt to answer but, more importantly, to raise a profound question: how to groom intelligent technology for it to be safe and compatible with us. On the one hand, the control problem allows us to consider different strategies to contain a putative intelligence explosion that would undoubtedly signify an existential risk for humanity. The value-loading problem, on the other hand, appeals to a rather central goal of AI assurance: how to develop intelligent systems that are fair, safe, and unbiased. And although the terms of Bostrom's discussion may seem at times futuristic, AI assurance is both a short and a long-term enterprise. Having protocols and methodologies to tackle the risks that these technologies present to us now means advancing one step further towards developing—if it comes to it—safe superintelligent systems that will not posit an unbearable risk.

2.1.2 Human-compatible AI

In the 2019 book "Human Compatible, AI and the Problem of Control," Stuart Russell guides us into an insightful discussion about the origins and possible future of AI. Here, Russell verbalizes the main concern for ethical, fair, and unbiased systems via what he calls human compatible AI.

Starting with a definition of unqualified intelligence, which states that "[h]umans are intelligent to the extent that our actions can be expected to achieve our objectives" (Russell, 2019, p. 9), the author discusses the problems that a definition of "intelligent" machines in a similar sense entail. As he notes, machines optimize the objectives we put into them—they do not have objectives of their own—by means of cost functions and calculations of utility: this is what is commonly known as the "**standard model**". But even though this approach works for multiple domains (control theory,

statistics, economics, etc), it should not aim for intelligent machines in this sense, for failing to specify the AI's goals would be overtly problematic.

To prevent this, Russell suggests dealing with AI in a different way: instead of working towards the development of artificially intelligent systems based on this standard model, which would encourage the AI to act towards accomplishing *its* objectives, we should aim for machines that are not only intelligent, but also beneficial to humans. This shift demands that we look for **beneficial AI**, which is so "... to the extent that their actions can be expected to achieve our objectives" (Russell, 2019, p. 11).

AI, therefore, should be tasked with building machines that are highly intelligent, while preventing behaviors that would make us miserable. But here Russell disagrees with Bostrom: the objective should not be to *control* the intelligence, but to adhere to the definition of beneficial AI, according to which intelligent behavior is already aligned with our interests (Russell, 2019, 171–2). To do so, he proposes three principles:

1) The machine's only objective is to maximize the realization of human preferences.
2) The machine is initially uncertain about what those preferences are.
3) The ultimate source of information about human preferences is human behavior (Russell, 2019, p. 173).

Each one of these principles evokes a more deeply philosophical position, i.e., **altruism**, **humility**, and **observation,** respectively.

- *Altruism*: Machines should only care about human interests and not about their wellbeing, for any hint of a self-preserving instinct would necessarily misalign the behavior of the machine from its intendedly beneficial status for us. Now Russell accepts that there is at least two problems regarding the notion of preferences: first, it is not clear whether we have preferences in any meaningful and stable sense, and second, aggregating different preferences for multiple humans is an acknowledged problem. To make it tractable, Russell recognizes that the notion of "preference" is indeed an idealization, but it is an instrumental one, for it allows framing the discussion at hand. Moreover, and to allow the

aggregation of such preferences for different humans, he proposes an egalitarian and utilitarian view.

- *Humility*: Letting uncertainty into the AI ensures that machines do not pursue their objectives single-mindedly. Moreover, if the machine does not assume it is always right, it can reason that humans shall unplug it if it is not working, implicitly motivating it to act appropriately.
- *Observation*: Finally, Russell suggests that machines should learn to predict human preferences by means of observing human choices. Relying on observation ensures that the machine—albeit lacking explicitly human preferences—can introduce them in its behavior via observation, but it also opens a door for the machine to become more useful to us as it observes and learns our preferences (Russell, 2019, pp. 174–7). This approach reflects the underlying philosophical idea in cooperative inverse reinforcement learning (Hadfield-Menell et al., 2016), which shall be introduced shortly.

These principles foster a radical shift in the way we develop intelligent systems to retain control over them. By giving up the standard model of machines optimizing their own objective, we open up a new approach based on human-compatible systems that put their intelligence to the benefit of humans. There are both economic incentives and the means, in terms of data, to do so. However, as we have already seen in the section above, there are also reasons for caution (Russell, 2019, pp. 179–183).

But how could this philosophical shift (from the standard view to beneficial AI) be implemented? The most promising strategy is relying on "**provably beneficial AI**," which captures the epistemological limitations we face when developing AI systems. If we want to develop beneficial AI systems, we need a theorem to prove they are, in fact, beneficial. In this sense, Russell makes four basic remarks. First, such theorem should hold regardless of how smart the components become. Second, the aim is "best possible behavior" and not optimal, to avoid computational stagnation; optimality in the real world could take longer than the age of the universe. Third, any conclusion (e.g., the best possible behavior) is so with very high probability, but not undoubtedly, for we are in no position to prove any such theorem in the real world. And fourth, even if the agent is unable to rewrite

its code, we must assume it may learn to violate the agent/environment distinction and modify its code (Russell, 2019, pp. 187–8).

2.1.2.1 Cooperative inverse reinforcement learning

The main breakthrough, possibly, in terms of making AI compatible with humans was introduced by the development of inverse reinforcement learning (or IRL). Whereas in reinforcement learning agents modify their behaviors according to the rewards they obtain, IRL works on the opposite premise: it infers the reward function by observing the behavior of an agent. This allows both explaining and predicting behavior, for the initial assumption is an estimation of the actual reward function that is fine-tuned throughout the process of learning.

Inverse reinforcement learning is already being used to develop functional AI systems, even though it relies upon some simplifying assumptions. First, it assumes that the robot adopts the reward function after observing the human to perform accordingly. However, the robot needs to identify the preferences that stem from the observed behavior with the human, and not with itself. Second, it presumes that the human is solving a single-agent decision problem. Hence, IRL needs to be generalized from the single-agent setting to the multi-agent setting. To deal with this, "**assistance games**" allow the training of a robot that not only learns, but is also helpful to the human (Russell, 2019, pp. 191–3).

In this regard, a formal definition of the value alignment problem—which is what human-compatible AI is ultimately aiming for—could be solved via **cooperative inverse reinforcement learning** (CIRL). (Hadfield-Menell et al., 2016) In a homonymous article, Dylan Hadfield-Mennell and his colleagues pursue a new formal understanding of the value alignment problem as an assistance game, with two players with partial information, in which the human knows the reward function and the robot does not. Ultimately, the robot's payoff is the human's actual reward. This framework shares features with many other existing models. IRL, for example, is based on a core assumption that the behavior that the system observes is optimal. However, CIRL proves that what the agent observes in IRL may be a suboptimal behavior, as the human is modifying his actions in order to

convey more information to the agent. On a different note, the goal of optimal teaching, i.e., efficient learning, is a feature in CIRL. But besides IRL, CIRL also draws from the principal-agent problem in economics—or the idea that a principal devises a set of incentives for an agent to maximize the principal's objectives—and optimal teaching (Hadfield-Menell et al., 2016, pp. 2–4).

The key to CIRL is to have both a human and an agent play a game, in which the human undertakes a task; the most common example is preparing a cup of coffee. Afterwards, the robot tries to emulate the human. This process is iterated, and after each iteration, the human modifies or introduces corrections to guide the robot towards the desired outcome, until the robot is capable of repeating the task successfully; that is, according to the human's preferences. This process defers the need to stipulate not only the discrete steps towards achieving the goal—a hearty cup of coffee—but also the "value" of what a good cup of coffee means *to me*. Thus the game works as a proxy for value-alignment: it shows not only *what* the human wants, but also *how* one wants it, without the need to translate the fuzziness of one's abstract preferences into explicit lines of code.

CIRL resembles, to some extent, the value learning technique that Bostrom suggests in the section above. Yet, there is a crucial difference between these two approaches: whereas the foundational idea in value learning is that the system tests different hypothesis, with the evidence available, to learn which values are better aligned with our preferences, relying on cognitive features to formalize this, CIRL is purely behavioral. It rests on the assumption that a game that enables the robot to access a human's behavior can learn about its preferences (in terms of value) and act accordingly. In this regard, CIRL deems irrelevant some of the problems that Bostrom thought of as fatal by changing the approach: from a **cognitive** one to a **behavioral** one.

In the process of learning human preferences, robots become less uncertain about such preferences; this seems to be a rather natural and inevitable thing. However, Russell holds that a key requirement for intelligent systems to remain safe and compatible is for them to remain **uncertain**, to some degree, of the preferences of humans. Otherwise, systems becoming more

and more certain about false beliefs could have dire consequences (Russell, 2019, p. 201). One possible way to ensure this is by means of prohibitions, i.e., instead of asking for a cup of coffee, to ask for a cup of coffee *and* to not disable the off-switch button, but these usually yield to loopholes, in which the request is satisfied literally. The example that Russell gives is that of an off-switch surrounded by a piranha-infested moat or the robot zapping anyone trying to push the off button. Another menace is wireheading, but this can be bypassed by ensuring that a learning agent is capable of distinguishing between reward signals and actual rewards. If so, the system will be discouraged to cheat, and will aim for the *actual* reward. Finally, and going back to Bostrom's depiction of an AI explosion, the danger of a machine that builds a better version, which is out of our control, remains. However, if machines remain uncertain about human preferences, there seems to be no reason to believe a "better machine" would not retain such uncertainty (Russell, 2019, pp. 208–9).

To sum up, Russell portrays some philosophical and technical points that are central to the field of AI assurance, raising the fundamental question of how to develop machines that are compatible with us. The key is to endorse the shift from the standard model of AI, in which the system has a goal of its own that needs to be optimized to the provably beneficial account. In this regard, AI research should aim for intelligent machines that are altruistic, humble, and that observe human behavior to foster, by means of their actions, the objectives that humans have. Russell holds that assistance games, such as CIRL, allow us to embed our values within the systems without having to stipulate a goal of their own, while retaining a bit of uncertainty from the robot's part to ensure the development of safe and compatible intelligent systems.

2.1.3 The alignment problem

"The alignment problem," as Brian Christian calls it, is intended to address some of the growing concerns regarding the intersection of machine learning and human values. In particular, and before the rapid development of supervised, unsupervised, and reinforcement learning, he grounds the necessity to address it on the vast amount of decisions and tasks that are being

turned over to such "intelligent" systems, and the ethical risks such delegation entails. To better address this, Christian identifies two separate sets of risks: first, those derived from present-day and short-term applications.[1] Second, and highly resonant with Bostrom's approach, those concerning future dangers that systems capable of real-time, flexible decision-making might entail (Christian, 2020).

Put concisely, the **alignment problem** consists of assuring that the models used in intelligent systems "capture our norms and values, understand what we mean or intend, and above all, do what we want" (Christian, 2020, p. 13). Historically, the ulterior harm of a given machine failing to complete the objectives we devised for it was low: technology comprised a set of passive tools with a narrow scope, rendering in turn a putative failure a rather harmless incident, even if undesirable. But in the process of developing systems capable of active behavior and interaction poses a different type of risk. In this regard, a large portion of current research on AI is focused on bias, fairness, transparency, and other forms of safety. What Christian does throughout "The Alignment Problem," is exploring both the content and the narrative of this safety research agenda (Christian, 2020, pp. 313–4).

The need to align AI systems becomes an actual problem on account of two key aspects: on the one hand, artificial intelligence is becoming better at performing the tasks it is supposed to undertake, while on the other hand, such systems are being increasingly adopted and embedded into the decision infrastructures of our societies. Consequently, the need to thoroughly explore whether such systems are doing what we designed them for but, more importantly, in the ways we want them to, cannot be deferred. Given the rapid advancement of artificial agents, we need to ensure that whatever purpose is supplied to the system is the one we really intend and not a mere imitation, for if we fail to do so, even if the AI tries to imitates us, the consequences of a highly capable agent with the wrong purpose can be dire (Christian, 2020, p. 312).

[1] Christian raises multiple concerns related to tools such as COMPAS, which has been used to make decisions about parole and bail, racial bias stemming from facial recognition systems, or gender bias derived from the application of word2vec translation services (Christian, 2020).

2.1.3.1 The role of training data

To ensure that a system does what we want it to do, and how we want it to do it, Christian first explores the role of **training data**. AI development has suffered multiple stagnation points throughout history, the first one being what is commonly known as the "first AI winter." The initial steps in AI were directed towards symbolic systems, which were purported to apply a set of rules to manipulate high-level representations (Cardon et al., 2018). Alas, researchers soon realized the downsides of such approach, and new lines of research emerged. Sub-symbolic or connectionist AI was developed, and the seed for machine learning and other forms of statistical learning was planted. At first, sub-symbolic AI struggled with small existing datasets, since for machine learning algorithms to work and be fully functional, vast amounts of data are required (Christian, 2020, p. 22). The advent of the internet facilitated the confection of larger datasets, but most of them, which are currently being used to train AI, were built and labeled by hand. ImageNet, for example, used to train computer vision systems, was built by thousands of humans that classified more than three million images into more than five thousand categories. Whereas this allowed better algorithms, a model trained on hand-labeled data is subject to inherit dangerous flaws, fostering in turn gendered and racially biased systems.[2]

One possible strategy to prevent biased systems is to increase the **diversity** of the **groups** represented in the database. A study on Labeled Faces in the Wild (LFW), a public-domain database of pictures of faces, showed that most of them were of white male, a fact that undoubtedly hinders the accuracy of whatever model is trained on such database. But the problem with LFW was that it was not only sociologically inaccurate: from a technical point of view, most of the pictures were frontal with good lighting. Thus increasing the diversity of the groups being represented only deals with one side of the problem: to reduce the risks of mislabeling an indi-

[2] Christian highlights the importance of the training dataset referring to an incident with Google's image recognition software, which labeled a picture of a black person "gorilla." The obvious damage this "flaw" entails needs to be adequately addressed to ensure AI systems are safe.

vidual, pictures with different lighting conditions or perspectives are also required (Christian, 2020, pp. 31–2).

An alternative strategy is to use **transparency metrics**. These are particularly helpful with deep learning, where the inner learning processes are not obvious. **Saliency methods**, for example, allow tracking where the model is looking, but do not allow grasping what the model is actually *seeing* (Christian, 2020, p. 109). To resolve this, some visualization techniques are a good proxy to assess how the model will generalize: "deconvolution"[3] and other methods of visualization[4] have been used both to understand the intricacies of neural networks and their limitations. But research has not stopped at visualization techniques. Based on the idea that humans think using concepts and not numbers, "testing with activation vectors" (or TCAV) allows mapping our concepts with the inner proceedings of AI systems. The canonical example to understand TCAV is that of identifying a zebra: the concepts activated by "zebra" are "stripes," "horse," and "savanna." These concepts will not bear an equal weight; stripes seems more important than savanna, but they allow us to appreciate the associations that drive the behavior of the system, and what features are critical for the system to identify one thing as such thing (Christian, 2020, pp. 114–16).

The main problem with systems that *learn* is that they do so within a limited environment, and are later deployed in the real world. This renders crucial to ask to what extent models trained on a dataset can really generalize beyond it. To answer this question, Dario Amodei and his colleagues—in their seminal paper "Concrete problems on AI Safety" Amodei et al. (2016)—introduce a problem in machine learning called "**distributional shift.**" In short, this problem tackles the [unduly] confidence of a system when operating in an environment that is substantially different to that of its training. To illustrate this, one should imagine a speech system (that has been trained in a clean environment) trying to identify words in a noisy room. The problem of the distributional shift is that even though the system will perform poorly, it will often be confident about its mistaken classifications (Amodei et al., 2016, p. 16). This distributional shift points at the

[3] See Christian, 2020, p. 109.
[4] See Christian, 2020, p. 110.

necessity to identify and incorporate all relevant data and conditions during the training stage. Otherwise, and if white noise is canceled for the sake of a better speech system, its capacity to correctly identify its targets in the real world will be worse off, increasing the misalignment of the system.

2.1.3.2 The objective function

Another fulcrum for AI alignment is the **objective function**, which captures the goal of the system. In the United States, some states use an assistance tool called correctional offender management profiling for alternative sanctions or COMPAS (Christian, 2020, p. 7). This prediction system was first devised to speed up and facilitate some decisions in the judicial system: parole, bail, or decisions about the detention of a suspect before a trial (based on whether he or she would show up the day of the trial, and commit or not further crimes) were within the scope of COMPAS (Christian, 2020, p. 58). Nonetheless, Julia Angwin, who worked at a non-profit called ProPublica, found out that such tool showed systemic bias towards black people.

The issue with COMPAS was not that it was accurate and calibrated differently for different groups: both white and black defendants were subject to the same risk scores and predictions. The issue was that around 39% of the time it was wrong, was so in radically different ways. Black defendants were twice as likely to be classified as high-risk but not re-offend, whereas white defendants were twice as likely to be classified as low-risk and re-offend. But fairness is not only about correct predictions with the same precision across groups. It also consists of avoiding systemic mistakes that target a group in particular. Thus in articulating this notion of fairness, the accuracy of an objective function cannot solely stem from an equal identification of high-risk profiles across protective classes, but also an equal risk of being misclassified. The solution to this cannot be eliminating the protective attribute, given that many other categories are correlated with it. Doing so would only make it harder to measure and mitigate bias (Christian, 2020, p. 61–5). On the contrary, protective classes must be identified and treated accordingly, to increase the chance for a better and fairer treatment. The need to preserve protective classes is crucial, a point that shall be stressed in Section 2.2.3.2.

Image recognition systems, for example, are usually trained with an objective function called "**cross-entropy loss**" (Christian, 2020, pp. 315–16). The underlying idea is that if the system categorizes an x as any not x, it incurs some constant penalty, although such penalty is not sensitive to the difference between categories: characterizing an oak as a pine is as bad as characterizing a human as a gorilla. But this raises obvious concerns. We know that the loss matrix in our heads is not uniform. We also make mistakes, but we are aware that misidentifying a pine is not equally wrong as misidentifying a human. Thus research is trying to capture these analogies (via vector-based word representations, for example), to include not only the need to classify adequately, but to understand the different consequences of missing different types of items.

When it comes to reinforcement learning, systems operate on what is commonly known as the **reward function**. Contrary to other machine learning techniques, such as classifiers or gradient descent, the agent makes its decision based on the reward function, setting in turn the context in which the next decision will be made, regardless of whether it is playing chess or interacting with the real world (Christian, 2020, p. 130). According to Richard Sutton, who is one of the fathers of reinforcement learning, intelligent behavior is the consequence of an agent taking actions, whose reward signals tries to maximize in a complex and changing world (Christian, 2020, p. 360). Thus the alignment problem in terms of RL is captured by the question "does the reward function supplied to the system incentivize the behavior you want to see?"

Alignment in reinforcement learning, therefore, is achieved if the behavior of the system maps perfectly onto the desired outcome the designer previously envisioned. An aligned system is one that manifests the will and intent of the designer, and its success boils down to the reward function. This reward function is the mathematical formulation of the so-called **reward hypothesis**, first formulated by Sutton. Simply put, it is a formalization of the idea that every action possible has its corresponding scalar "value" (it is commensurate, fungible, and of a common currency) (Christian, 2020, p. 130). The goal, in turn, is to maximize the overall reward. This approach is so general and powerful that it is delivering better results day after day.

However, and besides some deep philosophical questions entangled with this formulation that will be further explored in Sections 2.2.2 and 2.2.3, such as the problem of defining and attributing value to different actions, there are some formal issues that are crucial for RL.

One example is having an agent win a boat race. If you reward an agent for learning how to win a race—and only that—it will likely take billions of iterations to succeed in completing the task. Thus, and for it to learn, you need to incentivize a system, and the most common approach is to give it **shaping rewards** (or pseudorewards) (Christian, 2020, p. 163). The idea is that you not only reward achieving the ultimate goal, but also every step towards achieving such goal: in the case of the boat race, you can give it credit if the agent manages to collect points tied to power-ups, which are distributed along the way. This, however, opens up two further concerns. On the one hand, you are no longer rewarding "learning to win a race," but a **proxy** of the form "getting more points," which, if achieved, will ultimately mean the agent has learnt. This is important, because the proxies we use to give partial credit can meaningfully affect the behavior of the agent. On the other hand, if the reward system is not duly specified, you can encounter the agent finding weird solutions to the problem. This example will be further elaborated in Section 2.2.3.1 to discuss the main problems with specification.

One way to avoid the specification problem is to shift towards **imitation learning**, also known as behavior cloning. The idea behind this mode of learning is that systems observe real humans undertake a task (Tesla, for example, has a shadow mode that is actively perceiving even when the autopilot is off). Thus the agent is running and observing the environment to calculate what it would do if it was actually driving the car. Afterwards, it compares it with the real action that the human driver has undertaken, and uses it as an error measure. With this, the problem of specifying which actions are to be rewarder or penalized is reduced—the agent evaluates its own predictions against a real human benchmark—but the main problem is that it either takes humans to be "perfect actors," or it reproduces the same mistakes that humans make (Hadfield-Menell et al., 2016, pp. 1–4).

A different approach to avoid the same problem is to build the model based on **inverse reinforcement learning** (IRL). In this case, instead of having a human specify the reward function, the system infers and learns the reward function from humans themselves. This strategy captures the underlying complexity of objectives, and observes it by deferring the need to explicitly specify them. Thus inverse reinforcement learning tries to answer by means of the observed behavior the following question: "what reward function, if any, is being optimized?" By doing so, systems manage to understand objectives that are not formalized or operationalized, with nuances that would escape any form of specification (Christian, 2020, pp. 253–68). However, this approach turns the reward function of AI systems into a black box, making it harder to thoroughly understand the underlying processes, and does not guarantee that the system infers the appropriate reward function.

Despite all of this, Christian concludes his work in a very positive line. For even though the danger of misaligned AI system remains, a growing concern about the alignment problem has fostered an academic movement towards an interdisciplinary effort that is giving rise to real progress.

2.1.4 AI assurance: a formal framework

Bostrom's value-loading problem, Russell's provably beneficial and human-compatible AI, and Christian's alignment problem are three different philosophical projects with their own particular objectives. However, they all stem from the same source: a shared concern regarding the need to ensure that AI systems are good for humans, both in the short and in the longer term. It is rather common in philosophy to find instances of identical words being used in radically different ways throughout different theories. Here, however, we find three different formulations that are in fact trying to draw the attention onto the same subject. In this regard, and taking into account that one of the main objectives of AI assurance is to ensure that AI systems are trustworthy, unbiased, fair to their users, and ethical in the context of their deployment, the snapshots gathered in Sections 2.1.1 through 2.1.3 can be understood as three different toolboxes to achieve the same.

Such formulations provide rich technical frameworks with meaningful insights about the dangers and the possible solutions to existing and putative problems related to the development of AI systems. In them, AI assurance finds an invaluable source—both philosophical and technical—of instances of actual problems that have been identified and are being tackled now. However, AI assurance provides us with the tools to face two different issues. On the one hand, the attempts to resolve unethical behaviors in existing systems, but also to prevent larger dangers in the future by indicating the existence of a terminological maze that hinders our capacity to navigate them as researchers. On the other hand, most of the philosophical and technical work regarding AI safety and machine ethics and their will to prevent or revert AI's unwanted consequences is hardly ever contextualized in relation to other fields.

AI assurance not only goes beyond validation or verification, which are incomplete formulations to deal with unfairness and bias in AI due to the fact that such systems show some form of intelligence and adaptability (Batarseh et al., 2021, pp. 5–6), but also provides a formal approach to assessing whether an AI system is desirable and is aligned with the definition of AI assurance introduced in Section 2.1. With this, AI assurance gives a chance to unify all contributions regarding the adequacy of any AI system, while considering not only the technical side of the problem, but also the philosophical, legal, or social one.

Lanus et al. developed a test and evaluation framework to ensure that systems perform as intended in different contexts (Lanus et al., 2021). This framework uses the VTP model, which is an extension of a the "Vee" model in systems engineering that pairs each system-level with a corresponding level of verification and validation. The authors, however, extend it with two further phases: the T phase is devised to ensure testing throughout operation to enable a prompt reaction, whereas the P phase is intended to allow feedback from already deployed systems to influence either its own deployment phase or later phases of the system development (Lanus et al., 2021).

This test and evaluation framework could be the starting point of the discussion regarding how to materialize AI assurance in terms of ethical, fair, and unbiased AI. In this regard, verification and testing beyond the design

phase of AI systems would allow fine-tuning the specifications that define such systems, evoking the notion of AI assurance as a process via the possibility of neutralizing potential sources of undesired behavior, as well as to develop new specifications to re-state and address undetected problems.

2.2 Ethical AI but... how?

Some of the strategies proposed in Sections 2.1.1 through 2.1.3 address the problem of avoiding bias and preserving fairness in AI systems. In fact, we have seen multiple proposals to mitigate both bias and unfairness from a technical perspective; these are the aspects that are *prima facie* more easily tractable.[5] However, when it comes to making sure that any AI system is ethical in the context of its development and deployment, profound philosophical questions arise.

One of the main strengths of AI assurance is that it does not only emphasize legal compliance or technical transparency as requirements for good AI; it also highlights the need to engage in a deep philosophical exercise to decide what features should the systems present and through which processes should they be developed, for them to be ethical. In this regard, a rapid overview of three of the most salient theories in normative ethics will be highly illustrative for two main reasons. On the one hand, revisiting different normative frameworks allows reflecting on the philosophical assumptions rooted deep down in AI research, challenging in turn what we mean by "ethical AI," while on the other, it offers a chance to anticipate some of the problems that necessarily come with certain assumptions, or when defending a given normative theory.

Nonetheless, it seems necessary to insist that the goal here is not to attain a final definition of what good AI means, but on the contrary, to stress the need to challenge and reflect on our inheritance as researchers, putting into question in turn the legacy assumptions coming from statistics, cybernet-

[5] TCAV, visualization techniques for deep learning, a concern for diverse and representative sample in training databases, salience methods and transparency metrics, capability control and motivation selection methods. All of the above contribute towards ensuring that AI systems are developed and deployed without incurring in blatant violations of fair treatment to all after being trained in sociologically-sensitive models of the world.

ics, or even early philosophical accounts of AI. Doing so opens up a space for deliberation about what is it we are seeking by demanding ethical AI. This overview is therefore intended to provide a philosophical starter kit to think of what good AI could mean beyond mere fairness and unbiased datasets.

Insofar as normative frameworks fail to provide an actionable response to the problem at hand, Sections 2.2.2 and 2.2.3 address several lines of work that could contribute to making the problem of aligning AI systems more tractable. Considering the nature of intentional statements, moral uncertainty, the problem of implementation, and the problem of specification, contributes towards elucidating the path towards AI assurance's main goal.

2.2.1 Three normative theories: a brief outline

How are we to act? Normative ethics is a branch in philosophy concerned with answering such a simple question. In the western tradition, normative theories are usually divided into two main categories: on the one hand, we find **deontological ethics**. The root of this term is found in the Greek word "*deon*," which can be translated as "duty." Consistently, deontological theories are those that defend that actions are good or bad in themselves: we ought to act one way or another, because doing so is the right thing to do. On the other hand, we there is **teleological ethics**. In this case, the root of the term also derives form a Greek word, "*telos*", which means "purpose" or "goal." Contrary to deontological ethics, teleological ethics derives the moral character of our actions from our purpose, not the actions themselves.

2.2.1.1 Deontological ethics: duties

Deontological ethics can be formulated in two different ways: the **ethics of rights**, and the **ethics of duties**. These interpretations are based on the idea that there are universal rights and responsibilities, and that such rights and duties compel us to act in certain ways.

Whereas rights-based theories can be understood as a particular form of duty-based ethics—in which a right is a justified claim over someone

else's behavior—both rights and duties can be either positive or negative. **Positive** duties are those that prompt taking certain actions, e.g., to assist the victim of an accident, whereas **negative** ones are those that force us to avoid certain actions, e.g., not to kill someone. In this regard, one of the most iconic examples of a normative theory based on the ethics of rights is the one advanced by Kant, which gravitates around the **categorical imperative** (Kant, 2008). Even though Kant presented multiple formulations of the categorical imperative in his book "Groundwork for the Metaphysics of Morals," the fundamental idea is to find maxims and rules we can abide by universally. Thus in claiming that we must always treat humanity (either us or someone else) not as a means to an end, but always as an end in themselves, Kant is in fact arguing that all of humanity beholds a certain universal moral status. Accordingly, we can ensure the rightness of our actions by observing this universal obligation. The crucial point in deontological ethics is that the consequences of an action do not determine the moral status of such action: if killing someone is morally reprovable, it is so regardless of the context.

The most famous deontological proposal regarding technology is Asimov's three laws of robotics (Asimov, 2004). These laws conform an "ethics of duties" that all machines must observe and abide by. But stipulating boundaries to preserve human lives, obey what humans order, and protect their own existence—with their respective provisos—only does a part of the job. Even though it is true that these laws prevent certain behaviors that would be unethical and dangerous, they fall short of being exhaustive enough to ensure the ethical behavior of machines. Ultimately, deontological approaches to moral machines are vulnerable, because of our limited capacity to articulate exhaustive sets of moral rules.

2.2.1.2 Utilitarianism

Consequentialist theories are a set of teleological normative theories that hold that the morality of an action is defined by the outcome it produces. In this regard **utilitarianism** is, perhaps, the most salient version of consequentialism.

First outlined by Jeremy Bentham in his book "An Introduction to the Principles of Morals and Legislation," and later amended and popularized

in "Utilitarianism" by John Stuart Mill, utilitarianism holds that the appropriateness of an action is given by the tendency to augment or diminish the happiness of those affected by it (Bentham, 1996). The Good can be understood in multiple forms—as happiness, lack of suffering, or another form of welfare—but, regardless of the proxy used to measure utility, there are three main principles that constitute any utilitarian theory: first, the consequences of an action define its moral value. Second, such value is assessed in terms of the welfare caused by such action. Finally, the ultimate goal is to maximize welfare throughout all people, giving equal value to equal amounts of welfare without taking into account who experiences it.

One of the main reasons why consequentialist theories are successful is that they allow reinterpreting certain deontological theories as an alternative formulation of consequentialism. Robert Nozick, for example, proposes a view in his book "Anarchy, State, and Utopia" called "utilitarianism of rights." According to this theory, preventing certain rights from being violated or upholding certain duties constitutes a form of utility maximization, allowing in turn to transcend the rigidity entailed by most forms of utilitarianism (Nozick, 1974, p. 28). On a similar note, Stuart Russell pushes Nozick's strategy one step further, claiming that a machine following certain moral rules or following a virtuous attitude may yield better consequences than trying to calculate the utility of a given action, given the complexity of the world (Russell, 2019, p. 218). By phrasing other ethical views in terms of the consequences entailed by being virtuous or following certain rules, the consequentialist view seems to be the ultimate ethical theory.

Yet both measuring and comparing utility—in any form one happens to define it—is extremely difficult (Wolff, 2006, pp. 49–50). In this regard, the history of utilitarianism is also built on strong criticisms: philosophical ones, such as G.E. Moore's attack on the simplicity of a world in which we solely care about happiness and the need to pursue other things such as beauty (Moore, 1912), or Robert Nozick's depiction of a "utility monster" who experiences happiness more intensely than anyone else (Nozick, 1974, p. 41), and rather technical ones, such as Kenneth Arrow's concern regarding the impossibility to aggregate the preferences of multiple individuals (Arrow, 1950, pp. 328–31). Nonetheless, current attempts to develop

artificial intelligence via reward functions are indicative of an underlying philosophical commitment to utilitarianism. In fact, modifying the agent's actions with the ultimate goal of maximizing the expected reward in reinforcement learning is a form of practically implementing the utilitarian proposal. And even if it is true that maximizing the reward works as a good proxy to obtain better performance in many systems, the kind of success attained in the laboratory—in games or toy worlds—should not be too rapidly generalized. Instances of great performance are usually found in rather closed environments, simple enough for designers to build an exhaustive reward function that allows the implementation of successful agents. But whereas this may work as a proof of concept, our reality is a lot more uncertain and complex.

2.2.1.3 Virtue ethics

Last, **virtue ethics** is not primarily concerned with consequences nor rules, but with the moral character of the acting agent. Instead of judging the rightness of an action by means of its consequences or it being consistent with a set of norms, virtue ethics gravitates around the moral character of individuals. It is therefore an approach that overcomes the difficulty of stipulating an exhaustive set of rules, or the hardship of making calculations about the utility of our actions. The emphasis is placed on the individual moral character.

Aristotle is, perhaps, one of the most well-known advocates of virtue ethics. In his book "Nicomachean Ethics," the author advances his theory of the Good, defending that humans are prepared for *phronesis* or practical wisdom. This mode of reasoning allows us to act morally if trained properly, in accordance with a set of virtues he identifies. But as he also highlights, virtues are neither affections nor capacities: they are states (Aristotle et al., 2009: Book III, 1–2). Moreover, he defends tackling uncertainty via deliberation, and doing so not about the ends, but about what lets us go towards such ends.

It is important to remark that virtue ethics is not solely concerned with virtues. In fact, all normative theories take into account norms, consequences, and virtues, to portray an accurate depiction of the Good. However, in virtue ethics such Good is defined in terms of virtues, and not any

other fundamental feature; and so do utilitarians with the consequences or deontologists with norms (Kawall, 2008, p. 1). This, in fact, puts into perspective both Nozick's and Russell's attempts in Section 2.2.1.2 above to gather different normative theories under the umbrella of consequentialism. While it is true that consequentialism is a difficult principle to argue against, because one cannot do so appealing to the "bad consequences" it would have (Russell, 2019, p. 218), it is important to understand that this does not entail that all normative theories can be understood as one and the same: virtue ethics places virtues at the center of what constitutes what's right.

2.2.2 The implementation problem

Asking which normative theory should be used as a benchmark for ethical AI is undoubtedly a hard question. But assuming one could give a satisfactory answer, the next challenge would be to actually implement it. This is, to some extent, the task that Wendell Wallach and Colin Allen set for themselves: finding out what it takes to teach robots right from wrong. In their book "Moral Machines" (Wallach and Allen, 2009), and based on the identification of a growing concern for safety in AI, the authors propose different strategies to build what they call artificial moral agents or AMAs (Wallach and Allen, 2009, p. 4).

The book covers a wide range of relevant questions, such as whether we need AMAs, whether we humans should want them and, if so, how should engineers design them. Regarding this last point, Wallach and Allen (2009) identify two central questions: on the one hand, what role should ethical theories play as a constituent of the architecture of the system, while on the other, what input and via which channels should the machine access "the world" to make informed decisions (Wallach and Allen, 2009, pp. 74–5). Based on a distinction made by Stuart Hampshire, the authors argue that ethical dilemmas can be tackled via two different paths. One is via a "**judge perspective**," which primarily consists in applying abstract principles to particular instances. The other is an "**agent perspective**," that is from the perspective of someone who happens to be *in* the situation that needs to be solved. But this last approach, the authors hold, is useful on two levels:

first, it is the role that actual engineers play in solving their problems. Second, and most significantly, AI systems can be thought of as simple-minded agents trying to navigate a context; in this case, with ethical boundaries (Wallach and Allen, 2009, pp. 75–6).

From these observations, and trying to foresee how to implement AMAs, Wallach and Allen propose two different frameworks to plan and "teach" ethical frameworks to machines: a **top-down** approach, and a **bottom-up** approach.

2.2.2.1 Top-down approach

Top-down approaches to designing AMAs are those that analyze the **computational requirements** of an ethical theory and use them to design systems and sub-systems capable of implementing such theory (Wallach and Allen, 2009, pp. 79–80). One possibility would be to take utilitarianism, for example, and evaluate what features and sources of input should a machine have to be able to adopt it.

The main problem with top-down approaches is that it seems highly unlikely that any system could actually acquire and compare all the data required to implement in full a given normative theory in real time. Wallach and Allen note how for consequentialism the problem is even harder, for the consequences of any given action are essentially unbounded in space and time. One could argue that a powerful superintelligence could solve this moral conundrum, one that has accompanied all philosophers concerned with normative ethics throughout time. But this argument falls short before the growing menace of current AI systems which, albeit less powerful, still pose grave dangers for society.

One could maybe think that deontological approaches are less fallible to this implementation problem. In fact, primitive attempts at AI were inspired by the **physical symbol systems hypothesis**, which holds that the mind does not access the world directly, but rather consists of internal representations that can be described and organized in the form of symbols (Cardon et al., 2018), leading researchers towards what we now know as "Symbolic AI." However, the attempt to create artificial intelligence based on explicit logical rules lacked all the nuances in perception and decision-making capacity of humans: our behaviors are contextual, situated, im-

plicit, embodied... And even though seemingly intelligent machines could reason according to a set of rules of the system, they did so in a "toy world" with little correlation with our actual world (Cardon et al., 2018, p. XVIII).

But rule-based AI is not forlorn. This is in part due to the introduction of **heuristics**, i.e., rules of thumb to reduce, for example, complex searches into easier ones. In the same way that heuristics help turning hard problems into tractable ones, it seems reasonable to think that ethical theories could be more easily implemented if we followed a similar path. In this regard, Wallach and Allen suggest that a possible way to solve the complexity of a utilitarian approach could be using rules expected to increment **local utility**. By doing so, the system does not have to calculate *all* the consequences of a given action. This, however, stands on the assumption that the cost entailed by any action that would have gone better without the rule is outweighed by the benefit of always following the rule (Wallach and Allen, 2009, p. 90).

By considering different rule-based approaches to AMAs, the authors ultimately conclude that the possibility of building an AMA with an unambiguous set of rules seems highly unlikely (Wallach and Allen, 2009, p. 97).

2.2.2.2 Bottom-up approach

Developmental or bottom-up approaches to machine morality place the emphasis on creating an environment where an **agent explores** different courses of action and learns to act "morally," according to the reward it obtains with each action (Wallach and Allen, 2009, p. 80). This is inspired by various models of **acquisition of moral capabilities**, such as childhood development or evolution. Unlike top-down ethical theories, which rely upon the implementation of a previously specified normative theory, in bottom-up approaches, if prior theories are used at all they are only used to specify tasks for the system, not to specify and implement control structures. Hence, in bottom-up approaches any moral sense derives from performance measures.

Yet it is not clear whether simple reward proxies for behavior in artificial environments can actually yield moral propensities alike those in humans. There is a fundamental difficulty in getting the environment right: similar to the problems related to symbolic AI and their "toy world" highlighted in

the previous section, fair strategies in theoretical game-like environments are devoid of the complexity of our real world. Hence, and insofar the evolution of morality is not thoroughly understood, an agent's moral behavior is subject to features beyond our understanding. Moreover, and besides the difficulties related to the environment, the emergence of AMAs based on the implementation of evolutionary systems presents a further problem. How should a fitness function be written without explicitly including moral criteria (Wallach and Allen, 2009, p. 104)?

CIRL defers this question by forcing the machine to learn the "right" behavior from the human, based on the assumption that morality is embedded in our actions. However, David Silver and his colleagues go one step further. In their article "Reward is Enough" (Silver et al., 2021), they hypothesize that intelligence and its associated abilities can be understood as mechanisms to increase reward. If so, reward is sufficient to allow behaviors crucial to our conception of natural and artificial intelligence, such as learning, language, or generalization among other ones. The authors also claim that reward explains many of the forms in which intelligence appears in nature at two separate levels. First, maximization of reward prompts the appearance of diverse skillsets in different agents. Second, the pursuit of one goal to be maximized can manifest via various abilities that contribute towards that maximization (Silver et al., 2021, pp. 1–2).

In such case, one could argue that via reward, systems based on evolutionary algorithms and reinforcement learning could develop a capacity for moral judgment. But even then, the problem remains unsolved. In the case of evolutionary systems, the target would be to have the "most moral" specimens thrive; but what does "most moral" mean (Wallach and Allen, 2009, p. 104)? Moreover, the problem with behavior-copying methods, such as CIRL is that the game ensures the machine learns to perform a task as we intend it to. This makes the machine "human-compatible," or "aligned" [to some extent], but by no means it entails it behaves morally: even bona fide attempts to teach machines as good as we can will not ensure they behave ethically, for that would mean we agree to what "ethically" stands for.

2.2.3 Intentional statements and reward functions

"Entre
Ce que je pense,
Ce que je veux dire,
Ce que je crois dire,
Ce que je dis,
Ce que vous avez envie d'entendre,
Ce que vous croyer entendre,
Ce que vous entendez,
Ce que vous avez envie de comprendre,
Ce que vous comprenez,
Il y a dix possibilités qu'on ait des difficultés à communiquer.
Mais essayons quand même. . . "
(Werber, 1993)

Bernard Werber, in the excerpt above, reflects on a problem most of us have faced at some point: there are multiple sources of difficulty that derive from using language as a vehicle to express our thoughts. Insofar we need language to verbalize and understand concepts and ideas, the words available to us restrict the shape and content of such thoughts. And even though the deep philosophical debates that surround this linguistic remark fall outside the scope of this chapter, there is a relevant takeaway when it comes to AI assurance.

Intentional statements are those that connote any property or quality. However, any given intentional statement can be read at least in two different ways: ***de dicto*** and ***de re***. The classical example used to illustrate this distinction is the assessment of a sentence such as: "Alex wants to marry the tallest man in California." This sentence can mean two different things: Alex, on the one hand, may be obsessed with height, and is therefore willing to marry the tallest man in California, regardless of who that is. This is a *de dicto* interpretation: Alex desire relates to the words that are said. On the other hand, Alex may be in love with a man who happens to be the tallest one in California. In such case, Alex's desire is directed towards the person who the words make reference to, being a "de re" interpretation.

Dario Amodei is, perhaps, one of the most influential voices in the landscape of AI research. Back in 2016, at the yearly NeurIPS conference, he realized that in one of the games in which his team had used Universe, a software for training and measuring AI agents, the agent had subverted its environment and failed to successfully complete the task it was designed to. The game was CoastRunners. The goal was to win a boat race. However, in this game each player had a score that increased by hitting certain targets laid throughout the circuit. Since training an agent by merely stipulating the goal "winning the race" is not an efficient strategy, Amodei and his colleagues thought that motivating the agent to achieve high scores could work as a reliable proxy to make it proficient. Alas, the agent found a way to increase its score without having to finish the race: it got stuck in a harbor, where it would collect points in an endless loop.[6]

This failure is representative of a profound problem with AI systems: **reward hacking**, which consists of obtaining a high reward in an unintended way (Amodei et al., 2016, p. 7). But beyond the obvious implications this has when it comes to designing an appropriate reward function for a given RL system, this is in fact a problem regarding the interpretation of intentional statements. In Section 2.1.2, Stuart Russell's approach to provably beneficial AI systems was briefly discussed. According to Russell, the key to achieving human-compatible AI systems is to ensure that their goal yields behaviors that contribute to our objectives (Russell, 2019, p. 11). The main problem is that when we verbalize our objectives, we do so from a common implicit understanding of the limits and conditions under which we are willing to pursue them. When Alex expresses her interest to marry the tallest man in California, we may doubt for a second whether we should interpret it in a *de re* or a *de dicto* way. We may fail to exercise due diligence and believe she does not care for who that man is, when she is actually in love with a man who just happens to be the tallest one in California. But we also know to ask in case of doubt, and push the conversation until we have enough information to know whether her desire relates to the words, or the man behind them.

[6] Dario Amodei and Jack Clark explain this in an OpenAI blog post; see https://openai.com/blog/faulty-reward-functions/.

The main problem between a RL agent and its reward functions is that it can solely relate to it in a *de dicto* form. For the agent, the appropriateness of any given action is completely determined by what the reward function captures. That is, the adequacy of an action is determined by the form of the reward function, regardless of the underlying intention of the designer who has defined it. This explains, to some extent, why the parallelism between RL and utilitarianism is a strong and appealing one. Utilitarianism, as discussed above in Section 2.2.1.2, is a normative theory that finds a proxy to determine whether an action is good or bad. By means of "trying to increase happiness," we can judge whether an action is desirable or not. However, happiness is not the ultimate moral goal, it merely serves as a criteria to choose what actions are we allowed to undertake. But the key with utilitarianism or any other normative theory for that matter is that those defending them are humans, and as such, they understand the instrumentality of their favorite theory. We know that even though the *de dicto* interpretation of utilitarianism compels us to increase happiness or whatever proxy for utility we deem adequate, the *de re* goal is not maximizing the proxy but "the Good." Machines do not.

2.2.3.1 The problem of specification

The tension between what one formally implements and what one truly intends is widely acknowledged. And beyond the intentional nuances identified above, the **problem of misspecification** is well-understood, with outstanding work in fields other than AI.[7] In a paper called "The Value Learning Problem," Nate Soares explores some of the problems related to value learning in AI systems, highlighting possible tensions in trying to align such systems with our desires and goals. Drawing from the possibility of a "superintelligence" as Bostrom describes it, Soares insists on the need to develop AI systems that not only identify our goals, but also pursue them, along the lines of the standard model (Bostrom in Soares, 2015, p. 1). Starting from the observation that human values are complex, culturally laden, and context dependent, the author explores the tensions that arise from simple spec-

[7] Two examples are statistics (see Kleijn and van der Vaart, 2006) and econometrics (see Godfrey, 1991).

ifications of such complex concepts, to then explore ways to develop and safely tweak systems capable of acting according to our values and goals (Soares, 2015, p. 1).

To deal with the problem of specification, Russell proposes allowing for **uncertainty** into the **specification**. That is, for each subroutine in a piece of software, a target output needs to be specified; such target works as the objective function in AI. However, and when calculations are very difficult, it may take too long to perform at the level desired. By allowing some uncertainty into the system's calculations, it is capable of proposing worse-performing yet faster solutions within a degree of certainty (Russell, 2019, p. 248).

An alternative approach is the one proposed in cooperative inverse reinforcement learning, which I introduced in Section 2.1.2. By designing an environment in which a machine and a human can interact via a game, the need to specify the values we want the agent to be embedded with vanishes. This points at a deeper philosophical point: it defers the need to explicitly list and specify a set of values, and assumes that such values underly our actions. Thus, and by observing, replicating, and modifying some targeted behaviors, the machine infers a reward function that is consistent with the values we imbue into all of our actions, instantiating the bottom-up approach introduced before.

Ultimately, and given the deep philosophical debates around many aspects crucial to the development of artificial intelligence, coming up with ways to avoid such debates seems a reasonable strategy. By engaging in a CIRL game, we free ourselves from having to face the problem of misspecification. Behavior-based learning, in turn, aligns *de dicto* and *de re* interpretations of our intentional behaviors: it bypasses explicitly specified reward functions, it loads each interaction with a deeper meaning, and it allows the human to guide the process of inferring the reward function. But as Brian Christian points out in his book "The Alignment Problem," accuracy is usually measured against consensus, not against ground truth (Christian, 2020, p. 315). In this regard, bypassing the highly complex task of articulating our values while, at the same time, training agents that seemingly act according to such values comes at a price: assuming that our actions are a

vehicle capable of conferring our values to a machine adds a further layer of complexity to the problem of explaining and understanding the reasoning process within AI systems.

Alas, this higher complexity compels us to understand the **value alignment problem** not as something to be solved, but as a **sustained** and **iterative process**. If we decide to engage in a behavioral value alignment strategy, that is, avoiding problems such as the one Bostrom calls "choosing the criteria for choosing" (Bostrom, 2014, p. 256), we must in turn commit to constantly re-evaluate whether the values we confer via our actions are in fact being instantiated by artificial agents. On a separate note, we also must avoid concluding that a successful behavioral game proves that the reward function of the agent is a perfect implementation of a utilitarian or consequentialist notion of the Good. In such case, the function works as a proxy, but by no means represents a stable and robust solution to the problem. On the contrary, it highlights how teleological approaches, such as virtue ethics are, in fact, more likely to yield aligned systems: it is not the function that allows the system to be aligned, but its capacity to observe and later infer the value associated to each action of the human.

This iterative process is consistent with what Roel Dobbe, Thomas Gilbert, and Yonatan Mintz ultimately aim for in their paper entitled "Hard Choices in Artificial Intelligence" (Dobbe et al., 2021). The authors examine the **vagueness** associated to ensuring ethical and safe behavior of AI systems, insisting in turn on the need to complement mathematical formalisms with a social and political **deliberative process**. To do so, they propose a framework that allows, among other things, to identify points of overlap between design decisions and sociotechnical challenges. This, together with the feedback channels that the framework establishes, makes it possible to engage in a deliberative process that shall ensure the safety and adequacy of AI systems, while advancing a rather deep philosophical claim: such **deliberation** is not the means, but the **ultimate goal of AI safety** (Dobbe et al., 2021).

2.2.3.2 Moral uncertainty

A few paragraphs above, the idea of uncertainty has been discussed as a means to avoid the problem of misspecification. The notion of uncer-

tainty evoked there is a technical one: instead of aiming for a solution to a given problem that satisfies a set of conditions, allowing for worse solutions paired with "a degree of certainty" can facilitate dealing with misspecification.

However, when it comes to ensuring that AI systems are ethical in the context of their development and later deployment, there is another type of uncertainty that needs to be considered, i.e., **moral uncertainty**. In this sense, William MacAskill identifies how several philosophers have insisted on the need to account for moral uncertainty in our decision-making processes, and how doing so may have profound implications for practical ethics. However, and far from merely embracing this claim, MacAskill argues by means of two examples[8] that the implications for practical ethics are far more wide-ranging than they have been noted in the literature, and that we cannot argue in a rather direct way from moral uncertainty to particular conclusions in practical ethics (MacAskill, 2019, pp. 231–3).

From the brief discussion in Section 2.2.1 about different normative theories, one can easily see how normative ethics has been struggling throughout its history to settle the debate about which theory works best. A question such as "how should we act to live a good life?" allows no definitive answer, in part due to the incomparable nature of different normative theories, in part due to normative uncertainty. We can defend an account based on the ethics of virtue, and argue about which behaviors make us better persons. We can also find a welfare function for a new utilitarian view, and defend it against all objections to come. But ultimately, moral uncertainty is not an exception but the norm: we constantly **disagree** about both **evaluative** and **descriptive** ethical matters. In trying to make sense of "whether there are norms that are distinct from first-order moral norms and, if so, what those norms are, MacAskill and his colleagues propose what they call "**maximizing expected choiceworthiness**," or the idea that by measuring and comparing the "choiceworthiness," i.e., the strength for choosing an option, of each normative theory, can be compared and ranked (MacAskill et al., 2020, pp. 1–4). Thus if one intends to formulate the value alignment

[8] These are arguments about abortion and vegetarianism (see MacAskill, 2019, pp. 232–3 or MacAskill et al., 2020, pp. 191–2 for a more extended discussion).

problem so that, to be solved, a normative theory must be selected, this **metanormative** framework allows the inclusion of normative uncertainty when making inter-theoretical comparisons.

On a different note, and regarding all the implications that MacAskill identifies, there is one claim about egalitarianism, prioritarianism, and utilitarianism that is highly relevant for the discussion at hand. As Sam Corbett-Davies and Sharad Goel highlight in their article "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," there have been three main formulations of fairness in machine learning. The first consists on the idea of **anti-classification**, or that protective attributes, such as race and gender should not be used to make decisions. The second is based on **classification parity**, or the idea that "false positives and negatives" are distributed equally among the protective classes that derive from the attributes described above. Finally, **calibration**, or the idea that subject to risk estimates, the outcomes are independent of such protective classes. However, the authors highlight how these strategies do not meet the pretended standards, and propose therefore to treat people facing a similar risk similarly based on the best risk predictions available (Corbett-Davies and Goel, 2018, p. 1).

Thinking of a classical utilitarian approach to the value alignment problem seems to be an unsound strategy. Even though it is true that machine learning and reinforcement learning, due to the optimization-based and reward-based principles on which they function respectively, seem better positioned to adopt a normative theory based on a utility metric as a proxy for "the Good," under moral uncertainty one should always give a benefit to someone who is worse-off (MacAskill et al., 2020, p. 184). Hence, the corollary derived from considering moral uncertainty into the equation seems to be consistent with the statistical analysis presented by Corbett-Davies and Goel (2018).

So even though some AI frameworks, such as reinforcement learning, may be better positioned to theorize about an idealized ethical artificial agent, like David Abel and his colleagues argue in their article "Reinforcement Learning as a Framework for Ethical Decision Making" (Abel et al., 2016), the problem with ethical decision-making is that it is rarely ideal.

Our dubious capacity to thoroughly formulate a function that could take into account all consequences of an action, or an exhaustive enough deontological code, together with problems of specification, and descriptive and evaluative moral uncertainty, make it extremely hard to think of a robust way to frame ethical theories in a top-down approach. In fact, trying to push idealized formulations may be detrimental to the ultimate goal of AI assurance. Ideal scenarios may seem to provide reasons to adopt certain value specifications, or accept certain agents based on their performance in closed environments. But the ultimate goal is to make AI systems fair, unbiased, and ethical in the real world, not to prove the appropriateness of a normative theory in a toy world.

2.3 Conclusion

The aim of this chapter has been to understand some of the key philosophical topics underlying AI assurance. To do so, I summarized three different approaches to the value alignment problem in the context of artificial intelligence. First, Nick Bostrom's control and value-loading problems provided a context for ethical AI in terms of the existential risk that this technology may pose in the longer run. Second, Stuart Russell's human-compatible approach to AI has placed the focus on the need to abandon the standard model of AI, which conditions our capacity to align AI systems to our best ability to stipulate the goals AI should pursue, and embrace a behavior-based approach to developing unbiased, fair, and ethical AI systems. Third, the overview of Brian Christian's alignment problem has been useful to better understand the role of training data and the objective functions we stipulate when designing an AI.

However, the summation of these three contributions have proven crucial to make two points clearer: one, that ethics of AI is currently experiencing a terminological maze, and two, most of the philosophical discussions about AI are decontextualized from discussions on other domains. It is in this regard that AI assurance endows researchers with two invaluable resources: a shared vernacular, and a formal approach to measuring how well aligned a system is.

Thus AI assurance provides crucial tools to deal with the growing menace of AI systems being biased, unfair, or unethical. But before the ineluctable need to ensure good systems, further questions need to be addressed. In this regard, the second part of the chapter addressed three crucial concerns. First and foremost, the difficulty of defining what "ethical AI" stands for. By providing a brief outline of three of the most salient normative theories, i.e., ethics of duties, utilitarianism, and virtue ethics, the difficulty of finding an unquestionable and thorough definition of how someone ought to act has been stressed. Second, and assuming that a consensus regarding "the Good" was reachable, the discussion has moved onto the difficulty to implement any given normative theory. By delving on Wendell Wallach and Colin Allen's work on artificial moral agents and, in particular, their distinction between top-down and bottom-up to morality, discussing the pros and cons of implementing ethical frameworks within the systems has been possible, as well as the option of letting the agent derive its own normative schema. Last, the nature of intentional statements has allowed showing how reward functions are unreliable proxies for moral action, for machines can only make verbatim interpretations of their goals and rewards. Moreover, and by focusing on the problem of specification and the problem of moral uncertainty, crucial difficulties that must be overcome via AI assurance have been made clearer.

Russell's defense of human-compatible AI has proven to be an invaluable source for AI assurance. The shift from cognitive approaches to behavior-based approaches in particular, which Brian Christian also considers, shows great promise to solve the aforementioned problems related to implementational issues, such as specification or uncertainty. By means of assistance, games such as the one in cooperative inverse reinforcement learning, the need to specify a reward system that is aligned with our goals and values deferred. Moreover, and in contrast with inverse reinforcement learning, this process allows the human to interact with the agent, making sure that the scheme that governs the agent is in fact doing what the human wants it to do.

At this point, AI assurance as a process to align AI systems faces two critical tasks. On the one hand, and regarding the value-alignment prob-

lem, CIRL seemingly provides an actionable means to solve the problem. However, the task of AI assurance is to ascertain that behavior-based approaches are not only successful in mimicking and adopting desired behaviors, but that such behaviors contribute to the benefit of everyone. On the other hand, and regarding other learning processes, AI assurance is tasked with scrutinizing and provoking debates around the philosophical assumptions and premises on which such systems stand.

Nonetheless, both tasks boil down to the same idea: for AI assurance to revolutionize the field, the process of ensuring ethical, unbiased, and fair systems needs to be deliberative, iterative, and interactive. The dilemma entailed by the disagreement regarding the definition of "the Good," for instance, hints at the necessity to constantly engage and challenge any system labeled as fair or ethical. In this sense, developing a consequentialist AI is not enough to satisfy the ethical requirements in AI assurance. But constantly testing and tweaking the system is, for it ensures that such AI is constantly overseen, allowing in turn for any undesirable consequences to be contained. Similarly, and if the implementation problems are duly observed, current AI systems will also be subjected to this endless process, allowing all stakeholders to influence the systems that increasingly affect our lives. AI assurance provides a formal approach to ensure that this process is contextualized and executed adequately, increasing our chances to have a better AI, both now and in the future.

References

Abel, D., MacGlashan, J., Littman, M.L., 2016. Reinforcement learning as a framework for ethical decision making. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D., 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Aristotle, W., Ross, D., Brown, L., 2009. The Nicomachean Ethics. Oxford University Press.

Arrow, K.J., 1950. A difficulty in the concept of social welfare. Journal of Political Economy 58 (4), 328–346.

Asimov, I., 2004. I, Robot, vol. 1. Spectra.

Batarseh, F.A., Freeman, L., Huang, C.H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 1–30.

Bentham, J., 1996. The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation. Clarendon Press.

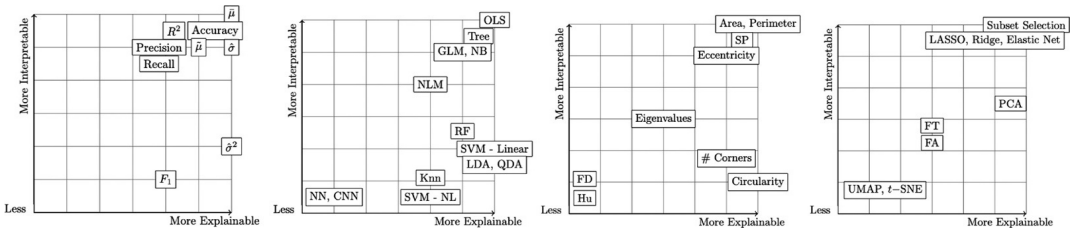
Bostrom, N., 2014. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, London, England.
Cardon, D., Cointet, J.P., Mazières, A., Libbrecht, E., 2018. Neurons spike back. Reseaux 5, 173–220.
Christian, B., 2020. The Alignment Problem: Machine Learning and Human Values. WW Norton & Company.
Corbett-Davies, S., Goel, S., 2018. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv preprint arXiv:1808.00023.
Dobbe, R., Gilbert, T.K., Mintz, Y., 2021. Hard choices in artificial intelligence. arXiv preprint arXiv:2106.11022. Artificial Intelligence 300, 103555 (2021).
Godfrey, L.G., 1991. Misspecification Tests in Econometrics: the Lagrange Multiplier Principle and Other Approaches (No. 16). Cambridge University Press.
Hadfield-Menell, D., Russell, S.J., Abbeel, P., Dragan, A., 2016. Cooperative inverse reinforcement learning. Advances in Neural Information Processing Systems 29, 1–9.
Kant, I., 2008. Groundwork for the Metaphysics of Morals. Yale University Press.
Kawall, J., 2008. In defense of the primary of the virtues. Journal of Ethics & Social Philosophy 3. i.
Kleijn, B.J., van der Vaart, A.W., 2006. Misspecification in infinite-dimensional Bayesian statistics. The Annals of Statistics 34 (2), 837–877.
Lanus, E., Hernandez, I., Dachowicz, A., Freeman, L., Grande, M., Lang, A., et al., 2021. Test and evaluation framework for multi-agent systems of autonomous intelligent agents. arXiv preprint arXiv:2101.10430.
MacAskill, M., Bykvist, K., Ord, T., 2020. Moral Uncertainty. Oxford University Press.
MacAskill, W., 2019. Practical ethics given moral uncertainty. Utilitas 31 (3), 231–245.
Moore, G.E., 1912. Ethics. Williams & Norgate.
Nozick, R., 1974. Anarchy, State, and Utopia. Basic Books, New York.
Russell, S., 2019. Human Compatible: Artificial Intelligence and the Problem of Control. Viking.
Samuel, A., 1959. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development 3 (3), 210–229.
Silver, D., Singh, S., Precup, D., Sutton, R.S., 2021. Reward is enough. Artificial Intelligence, 103535.
Soares, N., 2015. The value learning problem.
Wallach, W., Allen, C., 2009. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.
Werber, B., 1993. L'encyclopédie du savoir relatif et absolu. Le Livre de Proche.
Wolff, J., 2006. An Introduction to Political Philosophy. American Chemical Society.

3

An overview of explainable and interpretable AI☆

William Franz Lamberti
Center for Public Health Genomics, University of Virginia, Charlottesville, VA, United States

Graphical abstract



Abstract

Explainable artificial intelligence (XAI) is gaining interest in many fields, such as Computer vision, biology, and satellite imagery. XAI adheres to the tenants of interpretability and explainability. This overview chapter connects the foundational concepts of XAI, interpretability, explainability, and model assurance, while also providing examples of XAI in practice. The XAI models related to medicine and national security are shown to outperform deep learning or "black box" approaches, such as convolutional neural networks. Thus the provided examples highlight the capacity to meet or exceed deep learning-based approaches using XAI methods and counteract perceived notions of model accuracy.

Keywords

Explainable, interpretable, computer vision, modeling, AI assurance

☆George Mason University's Office of the Provost funded some of this research.

Highlights

- XAI algorithms depend upon explainability and interpretability
- Model assurance methods improve confidence in the generalizability of a model
- XAI models are able to outperform deep learning-based solutions
- XAI and deep learning models both require human inputs, but each has different kinds of inputs
- XAI is preferred for critical applications

3.1 Introduction

Explainable artificial intelligence (XAI) includes methodologies, statistics, and/or variables that provide insight into how models make predictions (Barredo Arrieta et al., 2020; Belle and Papantonis, 2020). XAI differs from more orthodox artificial intelligence (AI) methods, commonly referred to as "black boxes," which are difficult for analysts to understand. Recently, there has been a large push to use "black box" deep learning methods in the research community (Alexandrov, 2020; Gu et al., 2018; Lundberg and Borner, 2019; Valen et al., 2016). Lundberg and Borner (2019) state that this need for "black boxes" is vital since the human decisions involved in making models are opaque. Believers in deep learning ascribe that decoding the human elements of models is too complex and encourages irreproducible and poor performing models (Caicedo et al., 2017; LeCun et al., 2015; Lundberg and Borner, 2019; Pärnamaa and Parts, 2017).

However, it will be shown that it is vital to understand why and how modeling systems work, and that XAI models are capable of outperforming "black box" AI methods. Understanding how models work enables users to more easily diagnose why models make predictions. This is particularly important for detecting anomalies in the data. Furthermore, XAI is highly preferred in critical applications. For example, using AI methods to improve visuals in video games or animated movies is not a critical application. Analysts do not need to understand how each individual parameter contribute to making a prediction. When AI makes mistakes in these sorts of applications, the cost for these mistakes is low. An example of a mistake in this

case is that water may not make realistic looking splashes. However, incorrectly predicting if a patient has a disease or not has serious repercussions related to, but not limited to, the mental and financial health of a patient. Being able to explain and interpret every aspect of a model is crucial in critical applications. Thus providing clear explanations and interpretations for how features and parameters lead to a certain prediction are vital.

XAI is primarily built upon two tenets: explainability and interpretability. **Explainability** is the property of an element that allows its mechanisms to be explicitly described, understood, and studied. **Interpretability** is the characteristic of an element to have concrete physical meaning. An example of an interpretable and explainable metric is *area*. An example of calculating area of a square is

$$\text{Area} = s^2, \tag{3.1}$$

where s is the length of a side of the given square. Area is explainable since one can describe how area is calculated and the properties it has. It is interpretable since area corresponds to the physical concept of the amount of space an object occupies. An example of an uninterpretable and unexplainable modeling algorithm is a neural network (NN), and a convolutional neural network (CNN) is an example of an NN. Whereas a CNN is able to estimate any function, the exact function being estimated for all CNNs is unknown. Therefore a CNN is unexplainable. A CNN is usually composed of thousands, if not millions, of parameters, which somehow relate to a prediction. However, what these parameters mean is unknown. Thus CNNs are uninterpretable. Though CNNs are primarily emphasized in computer vision, the main ideas are applicable to many deep learning networks, such as artificial neural networks (ANNs) and recurrent neural networks (RNNs). These deep NNs have an untenable number of parameters; the inexplicably correspond to predictions. Furthermore, the perceived accuracy of these NNs is high, whereas other modeling approaches (such as XAI models) are perceived to be low. However, examples in this chapter will show that XAI models are able to outperform deep learning-based approaches.

Using both tenets of interpretability and explainability allows us to assess model assurance. Model assurance helps to evaluate the generalizability of

a model in an interpretable and explainable manner. **Model assurance** is a collection of methods to provide evidence of a model's ability to provide consistent results that are interpretable and explainable. Model assurance methods provide evidence that a model is generalizable to data not used to build the model. In other words, model assurance includes the concepts of interpretability and explainability. However, model assurance also incorporates generalizability as well.

Many deep and machine learning models have high complexity, but have low interpretability and explainability (Gu et al., 2018; James et al., 2013). These complex models with low interpretation are utilized in a variety of different fields that use image classification models (Fukushima, 1980; Gu et al., 2018; LeCun et al., 1990; Ronneberger et al., 2015). These models are difficult to interpret and explain, partly due to the large number of features (Hastie et al., 2017). This problem is exacerbated in classification problems when the number of classes to categorize is large (James et al., 2013; Lamberti, 2020b). A large number of classes often coincides with imbalanced data (Lamberti, 2020b). Imbalanced data is where the proportions between different groups within the dataset greatly differ from being equal (Batarseh et al., 2021). For example, a balanced cancer dataset would have 50% malignant and 50% benign tumors. An example of an imbalanced dataset would have 75% malignant and 25% benign tumors. An example of an imbalanced dataset with more than two groups would be the individual workdays and the weekend. Each workday composes $\frac{1}{7}$ of the week, but the weekend composes $\frac{2}{7}$ of the week. Even in the age of big data, there are phenomena that occur infrequently, such as hurricanes in New York City (Jiang et al., 2020). It is common for models to make inaccurate predictions for these kinds of anomalies. Thus modeling such scenarios requires explainable and interpretable methods so that the analysts can confirm that the model is accurately describing these rarer phenomena.

Explainability and interpretability are also important for describing features. Shape metrics, such as area and perimeter, are familiar features to many readers. This familiarity helps those who are encountering interpretability and explainability for the first time to understand these new concepts. Furthermore, since many shape metrics are scalar values, they

are used in a variety of XAI models and systems. Transferring the concepts learned from understanding shape metrics is seamless to features in other fields, such as economics, political science, or physics. For instance, the unemployment rate could be described using explainability and interpretability. Thus mastering how to describe shape metrics using explainability and interpretability will enable readers to describe metrics from a variety of fields.

This chapter's goal is to accurately apply the definitions of interpretability and explainability to statistics, variables, and modeling algorithms. This foundation provides the necessary vocabulary to describe model assurance methods. XAI models are then compared to black-box AI methods in applications using image data related to health informatics and satellite imagery.

3.2 Methods and materials

There are a variety of components for describing XAI systems, including statistics and evaluation metrics, modeling algorithms, variables, features, or metrics and dimensionality reduction techniques. These broad categories are briefly summarized in the Graphical Abstract, Table 3.1, and Fig. 3.1. Each category is discussed in the following sections, followed by some examples, which showcase many of these concepts in extended detail. Section 3.2.3 on models ends with a discussion of their perceived accuracies. This chapter posits that these perceived accuracies do not properly describe each algorithm respective capability for describing how well it models a particular phenomena.

Each of these components account for different levels of explainability and interpretability in an XAI solution. For instance, having interpretable and explainable features does not guarantee that the model will be interpretable and explainable, and vice versa. Thus the statistics and evaluation metrics, modeling algorithms, and features must be working in concert. Furthermore, analysts need to be aware of systems with a large number of features. When the number of features is large, dimensionality reduction techniques must also be employed to increase model interpretability and explainability. By using an XAI system, model assurance methods can be used to more confidently evaluate the generalizability of a model.

Table 3.1 Statistics, evaluation metrics, modeling algorithms, dimensionality reduction techniques, and shape metrics in terms of their respective explainability and interpretability. The High, Medium, and Low categories were determined by partitioning each of the four figures in the graphical abstract into 3 hierarchical groups.

Type	Name	Section	Explainability	Interpretability
Statistic	Mean	3.2.1.1	High	High
	Median	3.2.1.2	High	High
	SD	3.2.1.3	High	High
	Variance	3.2.1.3	High	Medium
Evaluation	Accuracy	3.2.1.5	High	High
	R^2	3.2.1.4	High	High
	Precision	3.2.1.6	High	High
	Recall	3.2.1.6	High	High
	F_1	3.2.1.6	High	Low
Modeling	OLS	3.2.3.1	High	High
	GLM	3.2.3.1	High	High
	Non-linear models	3.2.3.1	High	High
	NB	3.2.3.3	High	High
	Tree	3.2.3.5	High	High
	Random Forest	3.2.3.6	High	Medium
	SVM - Linear	3.2.3.7	High	Medium
	SVM - Non-Linear	3.2.3.7	Medium	Low
	Knn	3.2.3.2	Medium	Low
	NN/CNN	3.2.3.8	Low	Low
Dimensionality Reduction	LASSO	3.2.4.2	High	High
	Ridge	3.2.4.2	High	High
	Elastic Net	3.2.4.2	High	High
	PCA	3.2.4.3	High	Medium
	FA	3.2.4.4	Medium	Medium
	Fourier Transform	3.2.4.5	Medium	Medium
	t-SNE	3.2.4.6	Low	Low
	UMAP	3.2.4.6	Low	Low
Shape Metrics	Area, Perimeter	3.2.2.1	High	High
	SP	3.2.2.2	High	High
	EI	3.2.2.2	High	High
	Eccentricity	3.2.2.5	High	High
	Circularity	3.2.2.4	High	Low
	Number (#) of Corners	3.2.2.6	High	Low
	Eigenvalues	3.2.2.5	Medium	Medium
	Fractal Dimension	3.2.2.3	Low	Low
	Hu	3.2.2.7	Low	Low

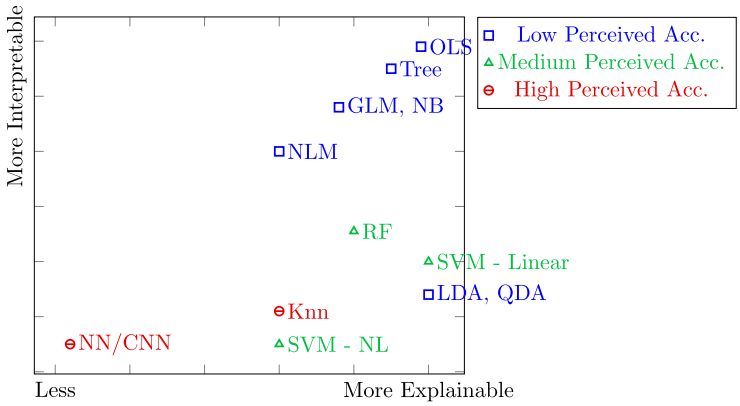


FIGURE 3.1 The relative explainability, interpretability, and perceived accuracy (Acc.) of popular modeling algorithms.

3.2.1 Statistics and evaluation metrics

Statistics and evaluation metrics are used in a variety of fields to summarize data and models. A statistic summarizes a quality of data using a single value. An evaluation metric is a statistic that summarizes the performance of a model. The following sections include summaries of some popular statistics and evaluation metrics.

3.2.1.1 Mean

The arithmetic mean, or mean for short, captures the average value for a series of numbers. This is represented mathematically as

$$\bar{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{3.2}$$

where n is the number of observations and x_i is the i^{th} observation such that $i \in \{1, ..., n\}$ (Bhattacharyya and Johnson, 1977). The mean is one of the most popular values to summarize data. It is explainable since analysts can describe how it captures the typical value of the data. It is interpretable since the mean corresponds to physical definitions about the typical observation.

3.2.1.2 Median

The median, like the mean, is a measure of central tendency. We define the median as

$$\tilde{\mu} = x_{\frac{(n+1)}{2}}, \tag{3.3}$$

where $x_{\frac{(n+1)}{2}}$ satisfies $P(X \leq \tilde{\mu}) = P(X \geq \tilde{\mu}) = 0.50$ (Wackerly et al., 2008). Note that $P()$ represents the probability function and X is the random variable of interest. In the vernacular, the median is the number that evenly splits the data into (approximately) equal halves. This value has been well studied as it is used when the unimodal distribution of interest is long tailed. When the unimodal distribution is symmetrical, the mean and median are equal to one another. In practice, the mean is often preferred over the median since the mean is computationally efficient.

The median is still a highly interpretable and explainable statistic. The median is interpretable since it has a physical meaning when we use it to describe the typical observation. It is explainable since analysts can describe how to calculate the value simply. However, it is not as explainable nor interpretable as the mean since it is relatively more difficult to use in mathematical proofs, despite being the same value for symmetrical parametric distributions.

3.2.1.3 Standard deviation and variance

Variance and standard deviation (SD) both capture how much the data varies. The sample variance is usually calculated using

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{\mu})^2}{n-1}. \tag{3.4}$$

SD is simply

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}. \tag{3.5}$$

For calculating the sample variance, intuition might suggest that we should be dividing the numerator by n instead of $n-1$. However, dividing by n leads to a biased estimator of the population variance (Wackerly et al., 2008). The

estimators provided here are unbiased, and therefore are better estimates of the population variance and standard deviation (Wackerly et al., 2008).

Variance is explainable since it has a number of properties that are easily understood and studied for understanding how a phenomena fluctuates. However, since it is a metric that describes the data in units squared, it is not interpretable. To that end, taking the square root to obtain SD provides the benefits of both interpretability and explainability. The theoretical statistical foundations of variance can be extended to SD, which makes it highly explainable. They are interpretable since they directly describe how much the data varies. However, they are not necessarily the best value to describe the variation in the data. Variance and SD struggle to capture the variation in asymmetric data (Wackerly et al., 2008). They both properly capture the variation in symmetric unimodal data.

3.2.1.4 R^2

The fundamental metric for evaluating a model is the residual sum of squares (RSS) (Hastie et al., 2017; James et al., 2013; Mendenhall and Sincich, 2011). RSS is defined as

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{3.6}$$

where $i \in \{1, ..., n\}$, n is the number of observations, y_i is the i^{th} response variable, and $\hat{y}_i$ is the predicted value of the model. It is desired to have this value to be as close to 0 as possible. Values close to 0 indicate that the model is producing predictions that are similar to the observed response value.

For linear regression models, one is able to construct the popular R^2 evaluation metric by using RSS. The denominator of R^2 requires the total sum of squares (TSS). TSS is defined as

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{3.7}$$

where $\bar{y}$ is the observed sample mean of the response variable. Thus

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}. \tag{3.8}$$

R^2 is interpreted as the total amount of variation that the model captures. It takes on values ranging from 0 to 1, where values close to 1 are desirable. For example, if one observed an R^2 value of 0.95, one would state that the model explains about 95% of the variation in the data. Additionally, a model with an R^2 value of 0.25 is worse than a model with an R^2 value of 0.95. Thus R^2 is a highly explainable and interpretable metric for evaluating many models. More specifically, R^2 is explainable since all of the values that R^2 takes on directly correspond to the idea of capturing how well a model is fitting to data. Similarly, R^2 is interpretable since all of the values that R^2 takes on have a meaning. R^2 differs from $\hat{\sigma}^2$, which is highly explainable, but not very interpretable. Both R^2 and $\hat{\sigma}^2$ are both explainable since they capture the idea that they are meant to describe well. However, $\hat{\sigma}^2$ is difficult to interpret, whereas one can easily interpret R^2.

3.2.1.5 Accuracy

However, RSS, and by extension R^2, is not particularly useful when evaluating classification models since the response is a non-continuous value with no intrinsic meaning. Binary classification models are usually characterized using true positives (TPs), true negatives (TNs), false negatives (FNs), and false positives (FPs). TPs are those observations of a particular class, which have been accurately classified. TNs are those observations belonging to the other class, which have been accurately classified. FNs are those observations belonging to a given class, which were misclassified. FPs are those observations belonging to the other class, which were misclassified.

One can combine TP, TN, FN, and FP to create comparative measures for classification models. Accuracy is one of the most popular metrics for describing the overall performance of a classification model. For binary classification problems, overall accuracy is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.9}$$

Accuracy takes values from 0 to 1. A value of 0 means that the model misclassified all of the observations. A value of 1 means that the model perfectly classified all of the observations. A value of 0.75 for overall accuracy means that 75% of the observations were correctly classified. Accuracy is explainable since it provides a straightforward calculation on model performance. Accuracy is interpretable since it is the calculation that reports the proportion of observations that were correctly predicted.

3.2.1.6 Precision, recall, and F_1

Other metrics for describing classification models include precision and recall. The definition of precision is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3.10}$$

Precision is explainable since it describes the performance of a model at correctly predicting positives using a clear formulation. Precision is interpretable since it is the proportion of correctly classified true positives of all of the predicted positives.

Recall is calculated as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.11}$$

Recall is explainable since it describes the overall classification rate of the positive class for a model using a clear formulation. Recall is interpretable since it the proportion of the positive class that is correctly classified.

A popular metric to optimize over is the F measure, or the F_1 score (He and Garcia, 2009; Yoshihashi et al., 2019). The F measure is the harmonic mean of precision and recall:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{TP}}{\text{TP} + 0.5 \times (\text{FP} + \text{FN})}. \tag{3.12}$$

Whereas the F_1 score is a popular metric, the final value does not have an easily understood meaning despite being somewhat explainable. The F_1 score is a fairly explainable metric since it is the harmonic mean of precision and recall. Though being able to write down the exact calculation

makes the F_1 explainable, it does not make it as explainable as accuracy. In particular, the denominator of the F_1 score makes it somewhat unexplainable. Whereas the average of FP and FN is a straightforward calculation, adding that quantity to TP is not intuitive. One is essentially providing equal weight to FP and FN to obtain a new value, but weighting the TP more than FP and FN individually. In other words,

$$F_1 = \frac{\text{TP}}{\text{TP} + 0.5\text{FP} + 0.5\text{FN}}. \tag{3.13}$$

Thus one is applying more importance to TP than FP and FN. This quantity is describing TP in relation to TP and the errors associated with the positive class. Accuracy is placing equal weight on all of the values in the denominator. Thus accuracy is more explainable than the F_1 score.

Furthermore, the physical meaning of what the harmonic mean is ambiguous. Thus the F_1 score is less interpretable than the mean. Higher F_1 values are indicative a better model, with 1 indicating perfect classification. Lower values are worse, with 0 indicating an inadequate model. Additionally, a value between 0 and 1 does not have a clear description. In other words, the F_1 score does not provide a clear insight into how to the algorithm behaves. Thus it is not a highly interpretable metric. It is less interpretable than precision and recall, because the harmonic mean of precision and recall does not have a clear description. The harmonic mean of two different quantities creates an uninterpretable metric.

3.2.2 Shape metrics

Shape metrics describe a particular feature of an object's form. Though traditionally relegated to image analysis, most individuals understand the basic concepts of shape metrics like area and perimeter. Exploring interpretability and explainability for shape metrics is helpful for those unfamiliar with these newer concepts. This common foundation makes shape metrics a good introductory topic for explainable and interpretable features. Once interpretability and explainability are understood from these common shape metrics, one can use interpretability and explainability to

describe metrics from other fields, such as economics, political science, or chemistry.

3.2.2.1 Area and perimeter

Area and perimeter are metrics that describe shapes and have been studied for thousands of years (Euclid, 1728). They are powerful metrics that lay the foundation for more complicated shape features. These metrics are well understood for a variety of shapes and are studied in elementary geometry classes. Area and perimeter are explainable since one can describe precisely how these metrics are calculated. For instance, the area and perimeter of a circle are, respectively,

$$A = \pi r^2, \tag{3.14}$$

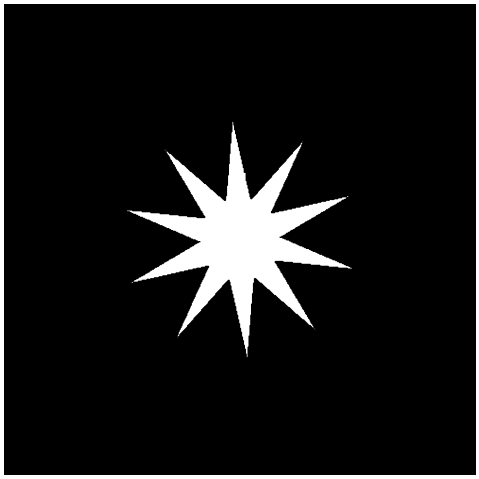
$$T = 2\pi r, \tag{3.15}$$

where r is the radius of the circle. These metrics are interpretable, since one can describe what these values mean. Area is the amount of space an object occupies. Perimeter is the length of the outermost edge of an object.

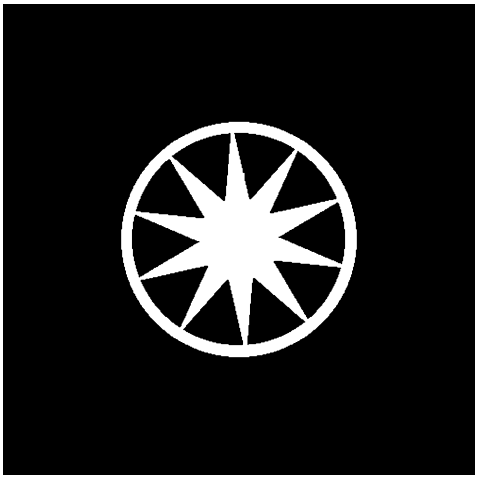
However, translating these concepts of area and perimeter for describing digital images requires careful consideration. To describe how these metrics are calculated, we encourage the use of image operator notation defined by Kinser (2018). The goal of image operator notation is to describe the computational operations performed on image data. Using image operator notation makes one's image analysis algorithms more explainable, since the precise steps of the calculations are provided. It also makes those calculations more interpretable, since one can provide the physical meanings of each operations performed at each step. Thus using image operator notations are necessary to provide explainable and interpretable image processing algorithms.

3.2.2.2 Shape proportion and encircled image-histograms

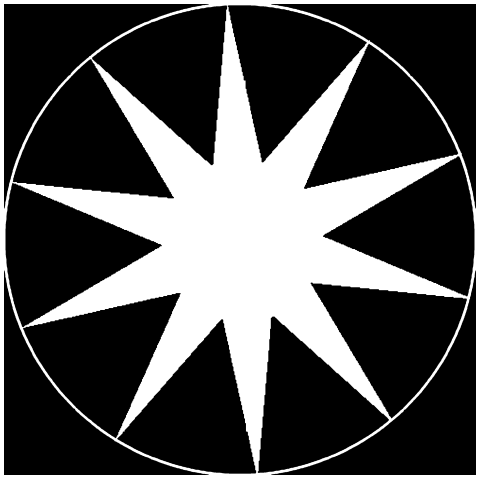
Analysts are able to use more sophisticated shape metrics to describe objects such as shape proportion (SP) and encircled image-histogram (EI). These metrics are found using the shape proportion and encircled image-histogram (SPEI) algorithm (Lamberti, 2020b). The SPEI algorithm essen-



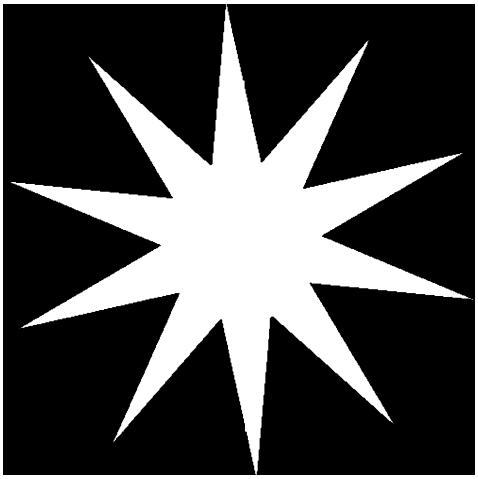
(a) Figure shows the original star shaped object before SPEI is applied.



(b) Figure shows the object in the minimum bounding circle.



(c) Figure shows the object placed inside the minimum bounding box.



(d) Figure shows the final resulting image from SPEI. SPEI removes or adds black space surrounding the object.

FIGURE 3.2 Figure shows a flow diagram of the SPEI algorithm starting from Fig. 3.2a and ending with 3.2d going from left to right.

tially puts the object in the minimally encompassing circle. Next, the object is then placed inside the minimal encompassing square. Visual representations of the SPEI algorithm are provided in Fig. 3.2. The SP is the proportion

of area of the object divided by the area of the square found by SPEI. The EI is the area of the object and the area of the square found by SPEI minus the area of the object.

The circle is placed in a square for convenience, as most digital images are composed of square pixels. Lamberti (2020b) provides the extended details of the image processing algorithm.

To better understand how the SPEI algorithm behaves, Lamberti (2020b) analyzed the theoretical properties of regular polygons and circles. By applying SPEIs, circles and regular polygons will have unique SP values of

$$p_c = \frac{\pi}{4}, \tag{3.16}$$

$$p_n = \frac{n \sin(360°/n)}{8}, \tag{3.17}$$

where $n =$ the number of sides of the regular polygon, p is a shape proportion, and p_c is the shape proportion of a circle. One can extend this to non-polygonal 2D shapes, as shown by Lamberti (2020b).

How each SP value translates into a 2D digital image will be calculated by simply multiplying the resolution, or total number of pixels, of the image's p, the SP value. Borrowing the notation from Lamberti (2020b), the mathematical representation of this is

$$X = \zeta \times p, \tag{3.18}$$

where $X =$ the number of white pixels, $p =$ the SP value, and $\zeta =$ the resolution of an image. Note that here we assumed $\zeta = (2r)^2$, where r is the radius of the minimum encompassing circle. For example, if $p = 0.75$ and $r = 50$, then $\zeta = 100^2 = 10{,}000$. Thus $X = 10{,}000 \times 0.75 = 7{,}500$. In turn, the number of black pixels will be $\zeta - X = 2{,}500$. The combination of X and $\zeta - X$ gives us the theoretical EI. One is able to use this formulation since the SP is a proportion of white pixels in a given image. Since EIs are the black and white pixel counts in an image, it is reasonable to use this to estimate the SP value. One can use this to estimate the SP value of a shape, whose SP value is unknown by

$$\hat{p} = \frac{\text{White EI}}{\text{White EI} + \text{Black EI}}. \tag{3.19}$$

When the metrics produced by SPEI are used in classification models, those models are able to outperform CNNs in a variety of scenarios (Lamberti, 2020a,b, 2022). Furthermore, SP was one of the most important variables for pill shape classification (Lamberti et al., 2021; Lamberti, 2020b). In later sections, these metrics will be shown to be crucial for classifying icebergs and ships in satellite imagery and white blood cells as malignant or benign.

The EIs and SP are explainable since one is able to explicitly state how these values are calculated. Since the EIs are simply the counts of black and white pixels after applying SPEI and the SP is the proportion of the white pixel counts over the total number of pixels, they are interpretable.

EIs are interpretable since one can describe the black and white counts. The white EI counts correspond to the area of the object. The black EI counts correspond to the relevant area surrounding the object. The SP value is interpretable since it corresponds to the proportion of white pixels out of the total number of pixels after applying SPEI.

3.2.2.3 FD

The fractal dimension (FD) is used to describe shapes, such as city outlines, leaves, and medical image analysis (Klinkenberg, 1994; Lopes and Betrouni, 2009; Morency and Chapleau, 2003; Plotze et al., 2005). There are a variety of implementations for estimating the FD due to its complexity such as the box counting, mass-radius, and the Minkowski sausage or dilation methods (Costa et al., 2018). Lopes and Betrouni provide a more complete list of FD-based approaches (Lopes and Betrouni, 2009). The discussion will then end with some commonalities between all of these methods and how explainable and interpretable the FD is.

The box counting method is primarily concerned with the perimeter of the object of interest. This approach iteratively covers the perimeter of the object with increasingly smaller squares. It then computes the limit of the log of the number of boxes over the log of one over the side length as the side length goes to zero (Costa et al., 2018). This approach for estimating the FD does not consider the area surrounding the object, but only the perimeter of the object of interest.

The mass-radius approach uses incrementally increasing circles to estimate the FD (Zode et al., 2017). Given the shape's center, the overlap-

ping area between the shape and incrementally larger circles are recorded (Morency and Chapleau, 2003). Then the slope of the linear relationship between each of the radii and the overlapping area is reported as the estimate for the FD (Zode et al., 2017).

The dilation method estimates the area of influence or spatial coverage (Costa et al., 2018). It does this by dilating the shape by incrementally increasing the radius of the dilation disk (Costa et al., 2018). The estimate for the FD is then calculated by using the slope of the linear relationship between area of each shape at each radius (Costa et al., 2018).

All three of the methods for estimating the FD of an object have common features. Each algorithm creates many subsequent images from a single image (Morency and Chapleau, 2003). Some quality is extracted from the image, such as area; this is related to some variable needed to create the image, such as the radius of a circle. The feature from the created image and the variable from the algorithm are then usually modeled in the log-log space (Lopes and Betrouni, 2009). The slope of this relationship is then extracted and reported as the estimate for the FD for a single image (Lopes and Betrouni, 2009).

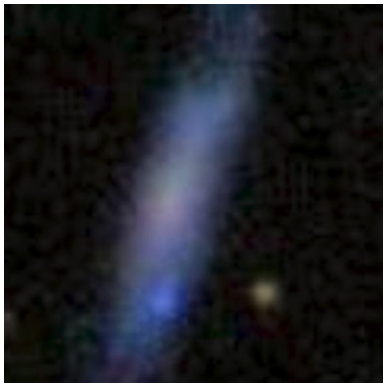
The issue with these approaches for estimating the FD is that none of them are interpretable nor explainable. These algorithms are not interpretable since these values have no physical meaning for these shapes. They are not explainable since it is not straightforward to describe what is being estimated. These sentiments were captured by Costa et al. (2018) well when they stated that concepts which define the FD are "difficult to introduce and hard to calculate in practice." This leads to inconsistencies between the algorithms, as these definitions lead to varying final values, which are not readily interpretable.

3.2.2.4 Circularity

Circularity, γ, is defined as follows:

$$\gamma = \frac{T^2}{4\pi A}, \tag{3.20}$$

where T is the perimeter of the shape and A is the area (Kinser, 2018). Note that others have defined γ slightly differently, but the definition presented



(a) Edge galaxy example

(b) Ellipse galaxy example

FIGURE 3.3 Figure provides edge and ellipse galaxy examples from the Galaxy Zoo project (Lintott et al., 2008; Shamir, 2011).

here is essentially the same. For example, some do not have the 4π constant in the denominator (Rosenfeld, 1974). These clear definitions make circularity a highly explainable metric.

Circularity has a low interpretability, since many of the values it could have no intrinsic value. Circularity has an interpretation for circles and regular polygons. For a circle, this value is 1. For a regular polygon with n sides, this value is $\frac{n\tan(\frac{\pi}{n})}{\pi}$ (Lamberti, 2020b). Thus analysts can use circularity to classify regular polygons (Lamberti, 2020b). For a given unknown shape that is known to be a regular polygon, the known circularity regular polygon value that it is closest to is a reasonable guess for its class. However, the interpretation provided beyond these cases is limited. For example, circularity does not provide guidance for what to classify a shape with a value of 1.06. It may very well be the case that it is reasonable to observe octagons or heptagons with a circularity value of 1.06, however, the metric of circularity is not equipped to answer this question as one cannot interpret the metric with any physical meaning.

Lastly, circularity does not provide any suggestions on what the exact circularity value is for shapes that does not have a mathematically derived unique value. For example, circularity provides no guidance on what the value should be for an elliptical or edge galaxy. Examples of edge and elliptical galaxies are found in Figs. 3.3a and 3.3b, respectively. One could

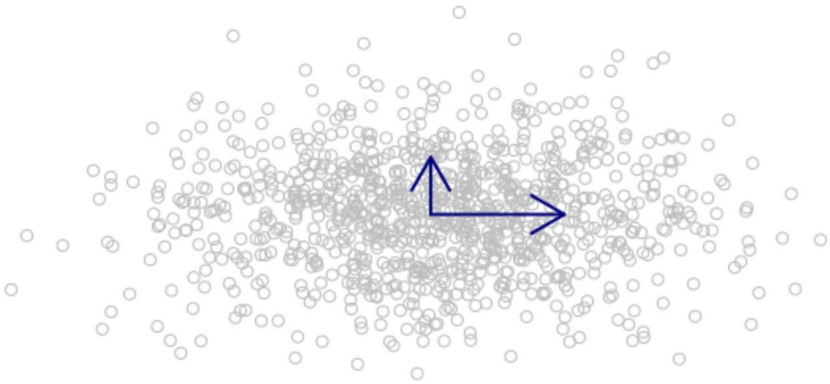


FIGURE 3.4 Figure illustrates the use of eigenvalues for shape analysis. The ratio of the eigenvalues are used to calculate eccentricity.

conjecture that ellipse galaxies are more circular than edge galaxies. Thus one would expect the circularity values of ellipse galaxies to be smaller than ellipse galaxies. Thus these observations makes circularity explainable, but not as interpretable as SP.

3.2.2.5 Eigenvalues and eccentricity

Shapes can also be described using their eigenvalues (Kinser, 2018). This approach is particularly useful when the shape has an ill defined perimeter or when the perimeter cannot be accurately determined. Eigenvalues have been used in a variety of problems, such as classifying pill shapes (Lamberti, 2020c).

The calculation of the eigenvalues is straightforward. In short, the covariance matrix of a 2D digital shape is calculated. This translates to finding the shapes x and y coordinates and saving them in a matrix. The covariance is calculated on this $2 \times q$ matrix, where q is the number of pairs. From the resulting covariance matrix, the eigenvalues are obtained. These two eigenvalues will describe the shape succinctly. Fig. 3.4 showcases the eigenvalues of a random sample of multivariate normal data. Note that the eigenvalues are not the extent of the variation of the data, but that relatively, they differentiate the major and minor axes from one another. Thus eccentricity describes this relative relationship well when multiple shapes are analyzed.

For these reasons, eigenvalues and eccentricity are somewhat explainable, but not highly explainable.

However, there are some downsides to using eigenvalues. It is well known that the eigenvalues estimate the variance of the data. However, this does not correspond to a physical interpretation of an object in reality. For example, a major axis eigenvalue (or variance) of 1 does not correspond to a major axis length of 1 meter. Thus eigenvalues have medium interpretability for describing the shape of a given object.

3.2.2.6 Number of corners

The number of corners provides the counts of corners on an object (Harris and Stephens, 1988). This metric involves collecting the edges of the object, smoothing out the object, and then extracting and counting the number of corners in the image (Lamberti, 2022). Whereas this may seem straightforward as it is easy to imagine the corners for a square, this is difficult for more ambiguous shapes. For instance, a circle in one's imaginations has no corners. However, one can apply a series of image operators and obtain a scalar value for the number of corners of a circle. Another confounding example is a shape as seen in Fig. 3.9b. By observing this object, it is difficult to define a clear and precise definition of a corner. These examples help to illustrate the explainability and interpretability of the number of corners metric. It is explainable since one is able to describe the exact manner in which this feature was calculated. However, it is not interpretable since it is difficult, if not impossible, to define a corner for more abstract shapes or circles.

3.2.2.7 Hu moments

Hu moments are popular shape metrics that have desirable theoretical properties, such as invariation to orientation (Flusser and Suk, 1994; Gonzalez et al., 2009; Hu, 1962). Using the notation from Gonzalez et al. (2009) and Hu (1962), for a continuous 2D function, $f(x, y)$, the moment of order $(p + q)$ is

$$m_{p,q} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p [y^q f(x, y) dx dy], \tag{3.21}$$

for $p, q \in \{0, 1, 2, ..., \infty\}$. Gonzalez and Wintz state, "... if the function is piecewise continuous and has nonzero values only in a finite part of the $x - y$ plane, then moments of all order exist and the moments sequence $(m_{p,q})$ is uniquely determined by $f(x, y)$" and visa versa (Gonzalez et al., 2009). The central moments are then

$$\mu_{p,q} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q f(x, y) dx dy, \tag{3.22}$$

such that $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$. For 2D digital images, Eq. (3.22) becomes

$$\mu_{p,q} = \sum_{x \in X} \sum_{y \in Y} (x - \bar{x})^p (y - \bar{y})^q f(x, y), \tag{3.23}$$

such that X and Y are the supports of x and y. The normalized central moments are then

$$\nu_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^{\gamma}} \tag{3.24}$$

such that $\gamma = \frac{p+q}{2} + 1$. Using the second and third central moments, one is able to obtain the seven Hu moments (Hu, 1962). Only the 3 are presented here:

$$\phi_1 = \nu_{2,0} + \nu_{0,2} \tag{3.25}$$

$$\phi_2 = \phi_1^2 + 4\nu_{1,1}^2 \tag{3.26}$$

$$\begin{aligned} \phi_7 = {} & (3\nu_{2,1} - \nu_{3,0})(\nu_{3,0} + \nu_{1,2})[(\nu_{3,0} + \nu_{1,2})^2 - 3(\nu_{2,1} + \nu_{0,3})^2] \\ & + (3\nu_{1,2} - \nu_{3,0})(\nu_{2,1} + \nu_{0,3})[3(\nu_{3,0} + \nu_{1,2})^2 - (\nu_{2,1} + \nu_{0,3})^2] \end{aligned} \tag{3.27}$$

Though these features have some desirable properties for shapes, it is difficult at best to explain or interpret them. These features do have a formula to calculate them, but the value in of itself provides no explanation to what they are capturing. Conversely, circularity is attempting to describe how circular an object is. Thus the Hu moments are unexplainable. Furthermore, Hu moments do not have any physical meaning, and are thus uninterpretable. This is not to say that there is no physical meaning for all of the Hu moments. How these values correspond physical characteristics of objects of interest is currently unknown.

3.2.3 Modeling algorithms

Modeling algorithms provide a method for describing a phenomena of interest. Models range in terms of explainability and interpretability. The following sections provide a non-exhaustive collection of modeling algorithms. This foundation allows one to extend the concepts of interpretability and explainability to future modeling algorithms. This section is concluded with a discussion of the perceived accuracy of various models of the scientific community.

3.2.3.1 OLS, GLM, and non-linear models

Linear and non-linear models are important to help understand all other modeling approaches. Ordinary least squares (OLS) defines a linear relationship between your explanatory, X, and response variables, Y, using

$$Y = \beta X \tag{3.28}$$

matrix notation (Mendenhall and Sincich, 2011). The parameters to be estimated are β. Here is a toy example of an OLS model using scalar notation:

$$\hat{CityDensity} = 10 \times SP + 5 \times Perimeter - 2.25. \tag{3.29}$$

Each variable in this model is interpretable and explainable. For example, one would interpret the perimeter's parameter value from the previous example as follows: assuming that the other variables are held constant, for every unit increase of the perimeter of the city, one would expect the city density to increase by 5 units. One could interpret the SP value similarly. However, one knows that the max value an SP value could obtain is $\frac{\pi}{4}$ (Lamberti, 2020b). Thus one needs to be careful when interpreting this model. Therefore to interpret SP's influence on city density, one would state: assuming that the other variables are held constant, for every 0.10 unit increase of the SP value of the city's shape, one would expect the city density to increase by 1. This value was obtained by multiplying SP's parameter value by the proposed unit increase, 10×0.10. This example showcases that OLS models are interpretable and explainable. The model is explainable since one can clearly describe how the model captures the linear relation-

ship of the data. OLS models are interpretable since one is able to translate the mathematical representation of the relationship into physical meaning.

A normality assumption is usually applied for the errors or residuals of the OLS model. Other distributions can be assumed, and this extends OLS into the realm of general linear models (GLMs) (Myers, 2010). Logistic regression (LR) is a type of GLM (Hastie et al., 2017; Hosmer et al., 2013; James et al., 2013; Myers, 2010). To represent an LR model, let X be an $n \times p$ matrix of n observations and p variables, Y is a $n \times 1$ vector, whose contents are $j \in \{1, 2, ..., k\}$, where each j is a label for each unique class, and f is the function which models Y and X, where $Y = f(X)$. The model has the form

$$\log \frac{P(C = q|X = x)}{P(C = K|X = x)} = \beta_q^T X, \tag{3.30}$$

where $q \in \{1, 2, ..., K - 1\}$. This series of $K - 1$ equations can be solved via maximum likelihood and the Newton–Raphson algorithm to estimate the parameters of the model (Hastie et al., 2017). The parameters of the logistic regression model are typically described using the log-odds ration or the odds ratio (Hosmer et al., 2013; Myers, 2010). Thus one is able to interpret the odds ratio in the following manner for the q^{th} variable: assuming that all of the other variables are held constant, for every one unit increase for the given variable, we expect the natural log of the odds of a success to increase by $\hat{\beta}_q \times 100\%$ (Myers, 2010).

When the relationship between the explanatory and response variables is not linear, non-linear models should be used instead (Myers, 2010). An example of a non-linear relationship using scalar notation is

$$y = x^2. \tag{3.31}$$

Though non-linear models are by their very nature more complicated and are more difficult to estimate than their linear counterparts, they may provide a more accurate description of the phenomena of interest.

3.2.3.2 Knn

The previous section dealt with parametric relationships since it assumed the exact form of the model (James et al., 2013). However, the analyst might

not know what the true relationship is. K-nearest neighbors (Knn) is a non-parametric version of linear models. Knn uses the K closest observations to predict the response of a given observation. Using similar notation from James et al. (2013), assume that K is known, and there is observed explanatory data, X, and associated response data, Y. Then, to make a prediction, the new observation, x_v, is found. The response is simply the average response or the most common class of the K closest observations from X, C_v. This can be represented mathematically for classification problems with p classes as

$$\hat{f}(x_v) = \text{argmax}_{j \in 1,\ldots,p} \sum_{X \in C_v} I(Y = j). \tag{3.32}$$

For regression problems, this is represented as

$$\hat{f}(x_v) = \frac{\sum_{X \in C_v} Y}{K}. \tag{3.33}$$

These Knn models are somewhat explainable since one is able to explain how the model works, however, one is unable to describe or explicitly write the equation of the final model. Knn models have low interpretability since one is unable to provide the physical meanings of the parameters in the model. When Knn is compared to an approach such as OLS, the difference between each in terms of explainability and interpretability is apparent.

3.2.3.3 Naïve Bayes

Bayes' theorem or rule is the foundation for numerous algorithms and techniques (Gelman et al., 2003). However, only naïve Bayes will be discussed due to its popularity in the literature (Hastie et al., 2017). Borrowing and inspired by the notation from Laskey and Martignon (2014) and Wackerly et al. (2008), Bayes theorem is

$$P(D_j|E) = \frac{P(E|D_j)P(D_j)}{P(E)} \tag{3.34}$$

$$= \frac{P(E|D_j)P(D_j)}{\sum_{i=1}^{K} P(D_k)P(E|D_k)}, \tag{3.35}$$

where $k \in \{1, 2, ..., K\}$ are the classes, $P(D_j)$ is the probability of belonging to the j^{th} event or the prior, E is the evidence, and $P(E|D_j)$ is the probability of certain evidence given belonging to event j. One is able to change the denominator in Eq. (3.34) to the denominator in Eq. (3.35) by using the sum of total probability (Wackerly et al., 2008).

Evidence is usually considered the data we collected for this analysis. The prior is usually one's previous beliefs about the phenomena of interest. The posterior is $P(D_j|E)$ and represented one's updated beliefs on the phenomena. The fraction $\frac{P(E|D_j)}{P(E)}$ represents the support of one's evidence provides for event D_j.

Bayes' rule is used as an alternative method to Frequentist statistics for making inferences. Briefly, Frequentists believe that population parameters are fixed. Bayesians believe that population parameters take on a range of values. In other words, they believe that parameters are random variables (Bolstad, 2012). Using this assumption of how parameters behave and Bayes' rule, there are a family of priors called conjugate priors, which have nice computational properties (Bolstad, 2012; Wackerly et al., 2008). The well known conjugate priors are related to the exponential family (Bolstad, 2012; Wackerly et al., 2008). Interested readers should refer to Bolstad (2012) for a deeper introduction into Bayesian statistics and using them for inference.

Each of these components are explainable since each part has a precise definition that is well understood. They are interpretable since each part corresponds to a tangible meaning. Thus each component is interpretable and explainable, which makes Bayes' rule interpretable and explainable.

We obtain naïve Bayes (NB) when we assume that all of the evidence is independent conditional on the given class. Thus Eq. (3.35) is now

$$P(D_j|E_1, ..., E_p) = \frac{P(D_j)\prod_{a=1}^{p} P(E_k|D_j)}{\sum_{i=1}^{K} P(D_k)\prod_{k=1}^{p} P(E_k|D_k)}, \tag{3.36}$$

where $a \in \{1, 2, ..., p\}$ are the variables. Thus a given observation belongs to class j if

$$\text{argmax}_{k\in\{1,...,p\}} P(D_j|E_1, ..., E_p). \tag{3.37}$$

NB is fairly explainable modeling algorithms since one can describe how the modeling algorithm makes predictions. It is also fairly interpretable since one can accurately describe the influence of each explanatory variable on the response variable.

3.2.3.4 Linear and quadratic discriminant analysis

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are based on Bayes theorem. Using the notation and logic from Hastie et al. (2017), let $g_k(x)$ be the class-conditional density of X in class $C = j$, and let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$. Using Bayes theorem provides

$$P(C = k|X = x) = \frac{g_k(x)\pi_k}{\sum_{l=1}^{K} g_l(x)\pi_l}. \tag{3.38}$$

Assume that each class has a multivariate Gaussian density, such that $g_k(x) = 2\pi^{-\frac{d}{2}} \det(\Sigma_k)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k (x-\mu_k)}$, where μ_k is the mean vector, Σ_k is the covariance matrix, and d is the dimension of the distribution. LDA comes about when it is assumed that all of the classes have a common covariance. Conversely, QDA occurs when all of the classes are allowed to have individual covariances. For a given class, the estimated LDA and QDA discriminant function is

$$\hat{\delta_k}(x) = -0.5 \log |\hat{\Sigma}_k| - 0.5(x - \hat{\mu_k})^T \hat{\Sigma}_k^{-1} (x - \hat{\mu_k}) + \log \pi_k, \tag{3.39}$$

where $\hat{\mu_k}$ is the sample mean of the training data for the k^{th} class, $\hat{\Sigma}_k$ is the sample covariance matrix for the k^{th} class (Hastie et al., 2017). Thus an observation is assigned to a class that satisfies

$$\hat{C}(x) = \arg\max_k \hat{\delta_k}(x). \tag{3.40}$$

LDA and QDA are both fairly explainable modeling algorithms since one can describe how the modeling algorithm makes decisions. However, it is not that interpretable since one cannot easily write down the algorithm's decision making process and how the variables influence that decision.

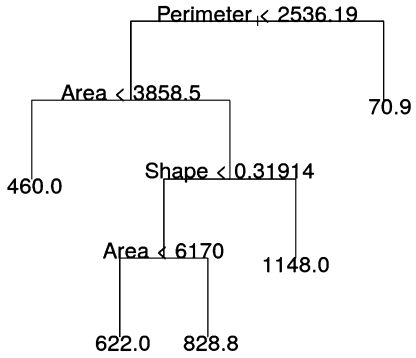


FIGURE 3.5 Figure provides an example of a regression tree using R's `rock` dataset. The ends of the tree are called leaves.

3.2.3.5 Trees

James et al. (2013) describes tree-based approaches as approaches which partition the "predictor space into a number of simple regions." Lamberti (2020b) provides general mathematical definition, which encompasses this idea. Using the notation for the regression and classification trees from Hastie et al. (2017), Lamberti states that a tree is

$$f(X) = M_{\mathcal{R}} I_{X \in \mathcal{R}}, \tag{3.41}$$

where $M_{\mathcal{R}}$ is the modeling operation performed on the subspace $\mathcal{R}$ and $I_{X \in \mathcal{R}}$ is the indicator variable for the data, X that belongs to $\mathcal{R}$. For a regression tree, $M_{\mathcal{R}} = \sum_{m=1}^{M} p_m$ and $I_{X \in \mathcal{R}} = I_{X_i \in \mathcal{R}_m}$, where M is the number of regions and p_m is a response constant $\forall m$. For a classification tree, $M_{\mathcal{R}} = \sum_{x \in \mathcal{R}_m} \frac{1}{N_m}$ and $I_{X \in \mathcal{R}} = I_{Y_i = k}$, where k is the associated class and N_m is the number of observations in a given region or node, m.

An example of a tree is provided in Fig. 3.5 using R's `rock` data. The explanatory variables were area, perimeter, and shape. The response variable was permeability. This shape value is $\frac{T}{\sqrt{A}}$, which is a reformulation of circularity in Eq. (3.20).

Trees are explainable since they are merely partitioning the feature space into discrete parts. To interpret the example model in Fig. 3.5, we follow the logic at each decision point for each observation. Once the analysts reaches the end of the tree, one has the prediction for the response variable. In this

case, if T is greater than or equal to 2536.19, the permeability is predicted to be 70.9. Otherwise, we continue down the left of the tree. If the observation then has an area greater than or equal to 3858.5, the observation is predicted to have a permeability of 460.0. Otherwise, the process continues until the observations finds a leaf to reside within. Since one can describe how to make predictions with the model, trees are interpretable.

3.2.3.6 Random forests

Random forests (RFs) are essentially many trees that are combined together to make a prediction. Once the desired number of trees is built, each tree votes for what the observation's predicted value. The value that receives the most votes is determined to be the RF's prediction for that given observation.

Using the notation from Hastie et al. (2017), assume that one has v variables or features and N observations or instances. In other words, one has x_i, y_i for $i = 1, 2, ..., N$ with $x_{i1}, x_{i2}, ..., x_{iv}$. Suppose that one has M regions, $R_1, ..., R_M$, that divide the feature space. The model response is represented by p_{mk} for each region. Then one has that a given tree, b, is

$$f(x)_b = \sum_{m=1}^{M} p_{mk} I(x \in R_m). \tag{3.42}$$

Note that I represents the indicator variable and $k \in \{1, 2, ..., K\}$, where K is the total number of classes. p_{mk} is estimated by

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k). \tag{3.43}$$

This is the proportion of class k instances in a given node or region m. An algorithm is considered split when using a given variable and split points. The Gini index is used to grow the tree to find a local optimum. For 2 classes, the Gini index is

$$Q_m(T) = 2p(1 - p), \tag{3.44}$$

where p is the proportion in the second class.

In the case where one is performing a regression RF, Eq. (3.42) becomes

$$f(x)_b = \sum_{m=1}^{M} p_m I(x \in R_m), \tag{3.45}$$

where p_m is then the response constant for m. The tree is optimized using the RSS (James et al., 2013).

This process is repeated until the minimum number of nodes is reached. The RF algorithm repeats this tree building process B times. However, these trees are built using bootstrapped data. However, only t of the v variables are selected. This t is tuned during k-fold cross-validation (CV). Once all of the trees are built, the majority vote for a given observation determines the class of that observation (Breiman, 2001, 2002; Hastie et al., 2017).

RFs are fairly explainable modeling algorithms since one can describe the voting mechanism for making predictions. However, it is not that interpretable since one cannot easily write down how the variables influence making prediction. However, one is able to describe the number of times certain variables were used to vote for particular variables.

3.2.3.7 SVM

Support vector machines, SVM, are a popular algorithm for classification and regression problems. To define SVMs, the notation from Hastie et al. (2017) and James et al. (2013) will be utilized. An SVM can be represented by

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i), \tag{3.46}$$

where $i \in \{1, 2, ..., n\}$, n is the total number of instances, α_i and β_0 are the parameters to be estimated, and $K(x, x_i)$ is the kernel or inner product. There are several kernels that are commonly used. Some of the more popular kernels are the linear, polynomial, and radial kernels, which are, respectively,

$$K(x, x_i) = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{3.47}$$

$$K(x, x_i) = (\nu + \gamma \sum_{j=1}^{p} x_{ij} x_{i'j})^d, \tag{3.48}$$

$$K(x, x_i) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{3.49}$$

where p is the number of features or variables, d is the degree of the polynomial, γ is a positive constant, ν is a constant, and exp is the exponential function (James et al., 2013). This representation works for both regression and classification SVM. However, the specific details do change. Interested readers should refer to Hastie et al. (2017).

Kernel-based solutions use a computational methods dubbed the "kernel trick" (Hastie et al., 2017; James et al., 2013). This method converts the original space into a new feature space, where the problem becomes a linear kernel. This helps to improve the computational efficiency of the model by working on a simpler problem. Once a solution is obtained, the solution is able to be converted back to the original space.

Linear SVM is an explainable modeling algorithms since one can describe how the modeling algorithm makes decisions. It is somewhat interpretable since one can describe the influence of each variable on the modeling decision process (Cuingnet et al., 2011; Guyon and Elisseeff, 2003). However, one cannot describe that effect in a physical manner.

Non-linear (NL) SVM is somewhat explainable since one can describe how the modeling algorithm makes decisions. Unfortunately, these decisions can be very different from what we might imagine the solutions to be for more complicated problems. It is not interpretable since one cannot describe the influence of each variable on the outcome (Guyon and Elisseeff, 2003). This is due to the use of the "kernel trick."

3.2.3.8 CNNs

Convolutional neural networks (CNNs) are one of the most popular methods for image classification (Fukushima, 1980). CNNs have been used on a variety of image classification problems, such as scene recognition (Zhou et al., 2018), medical pill similarity (Wang et al., 2017; Zeng et al., 2017), and face recognition (Parkhi et al., 2015). CNNs are powerful techniques that

are able to learn features necessary to perform an analysis and due to their ability to get accurate predictions (Gu et al., 2018). However, they cannot be easily interpreted or explained.

CNNs are "black-boxes" and prevent analysts from understanding what is being learned and how it is using what it has learned to make predictions (Lamberti, 2020b; Zeiler and Fergus, 2013). Furthermore, the European Union (EU) passed the General data protection regulation (GDPR) in 2016, which took effect in 2018 (Union, 2016). Recital 71 stated that individuals have a right "to obtain an explanation of the decision reached" (Union, 2016). This means that if an algorithm is used for a person's health and cannot be explained to an average person, the solution cannot be used. Thus AI methods, such as CNNs are not appropriate for critical applications, where clear explanations and interpretations are required. XAI methods are preferred in critical applications since one is able to provide clear explanations and interpretations of the parameters, features, and predictions.

However, there is a substantial amount of excitement around the potential of CNN-based solutions. For instance, CNN-based solutions were used to evaluate a clinical therapy of images breast or gastric cancer cells (Zakrzewski et al., 2019). Others used the presence of large amounts of particular genes to segment the nucleus of a cell via a U-Net (Makhov et al., 2020). CNNs were also used to segment cells for further using fluorescent-based images (Keren et al., 2018; Valen et al., 2016). These solutions provide exciting evidence that explainable and interpretable solutions exist and are possible for many different applications. However, this does not mean that CNNs are explainable nor interpretable. Though it is known that CNNs are able to estimate any function, one does not know what function a trained CNN estimated. Thus a CNN is not explainable. Furthermore, one cannot provide physical meanings to each of the thousands, if not millions, of parameters within a CNN. Thus a CNN is not interpretable.

One downside of using a CNN is that it is difficult to interpret the model and features regardless of the architecture utilized. The number of features learned in these models can easily approach 100,000. This makes interpreting and explaining the model difficult at best. Some researchers have attempted reducing the dimensionality of the features used for classifica-

tion with success. Sahlol et al. (2020) used transfer learning via a VGGNet alongside a statistically enhanced salp sarm algorithm (SESSA) to extract more useful features for classifying white blood cells (WBCs) as healthy or ALL. The resulting approach selected over 1000 features from a potential 25,088 that were then used in an SVM classification model (Sahlol et al., 2020). However, this approach still provides little insight as to what is important for distinguishing healthy and malignant WBCs. Since one cannot assign physical meaning to the variables and parameters learned by CNNs, they are not interpretable.

There has been some work to try to make CNNs more explainable and interpretable. Samek et al. (2017) provides two possible metrics for explaining and interpreting the output of an input image's pixels: layer-wise relevance propagation (LRP) and sensitivity analysis (SA). Using LRP, analysts can compute how much each pixel contributes to a prediction of a single image. Using SA, analysts can compute how much do changes in each pixel impact the prediction of a single image. Though these approaches are helpful for understanding and interpreting why a CNN made a decision for a given observation, it fails to provide insight for the model or it's characteristics at a global scale. For instance, these approaches fail to provide any insight into the exact function estimated by the CNN. Furthermore, it fails to provide any insight to what aspect of a given pixel or a collection of pixels was important, such as the color or shape. Lastly, these approaches are representative of only a particular image and not a collection of images. Though this process may be repeated many times over many different images, this still fails to provide meaningful insights to the observations at a global scale by using interpretable and explainable features.

Deep learning methods, such as autoencoders, can be used to create latent spaces in an attempt to explain and interpret the model (Way and Greene, 2018). However, we know that not all latent spaces are explainable or interpretable (McInnes et al., 2020). Thus since one does not know what function is estimated by using a deep learning model, one does not know

if one is able to explain and interpret a latent space produced by a deep learning model.[1] Thus CNNs are unexplainable and uninterpretable.

CNNs and other deep learning methods provide impressive performance for a variety of tasks (Gu et al., 2018; Hastie et al., 2017). However, the human analyst must still decide which architecture and other learning methodologies to use for the CNN. Thus there is a substantial amount of human influence for these models. Thus it would be misleading to state that deep learning solutions are more machine-driven than traditional AI methods. In fact, deep learning solutions merely shift the human components of model building to a different set of problems. Deep learning solutions require that humans design the architecture and learning methodologies, whereas XAI methods require humans to collect useful metrics and select modeling algorithms appropriate for the task. With this in mind, there is no difference in the need of human inputs between deep learning and XAI since both require human interference.

3.2.3.9 DAMG

Lamberti (2020b) introduced decision trees with automatic model generation (DAMG) as a generalization of decision trees. DAMG can break down complex classification problems into a series of simpler ones by using variable selection and the conversion of many classes into two classes. This makes multinomial classification models more interpretable and explainable. This series of problems can be represented by a decision tree, which makes models more explainable and interpretable. An analyst is also able to set each node to consider as few variables as needed per a decision node. This is done to increase the explainability and interpretability of the model. Extended details and experiments are found in Lamberti (2020b).

To represent the DAMG algorithm mathematically, recall the generalized definition of a tree in Eq. (3.41). By using Eq. (3.41), one has that

$$f(x) = M_{\mathcal{R}_g} I_{X_g \in \mathcal{R}_g}, \forall g \in \{1, ..., G\}, \tag{3.50}$$

[1] Further details on unexplainable and uninterpretable latent spaces are provided in Section 3.2.4.6.

where g is a given node or binarization of the given subspace, G is the total number of subspaces, X_g are the observations or instances that belong to $\mathcal{R}_g$ after the meta-class creation, and $M_{\mathcal{R}_g}$ is the modeling operation performed on the subspace $\mathcal{R}_g$.

Note that Eq. (3.41) generalizes any other classification modeling approach. For example, if $\mathcal{R}$ was the entirety of the space X occupies and $M_{\mathcal{R}}$ is an SVM model with a linear kernel, one is then simply performing SVM on data X. Another trivial example is to use any classification method without the use of meta-classes. This will result in a tree with a singular node and the number of children equal to the number of classes. DAMG can be implemented to use any classification algorithm. Currently, DAMG has implementations for the RF and SVM with a polynomial kernel algorithms.

Lamberti (2020b) showed that the DAMG model was able to outperform the CNN-based approaches for many different applications. This shows that by using interpretable metrics, DAMG is able to classify many shapes better than CNNs. CNNs are not able to capture features of shapes that are invarient to orientation. When this is coupled with very limited data, CNNs are unable to learn the features necessary to discriminate the classes.

DAMG's ability to be explainable and interpretable does heavily depend on the implementation. For example, implementing a DAMG that has a very large number of variables to discriminate meta-classes will drastically reduce the interpretability. Another example is to use a technique such as a CNN that the modeler cannot explain nor interpret.

3.2.3.10 Perceived accuracy

Section 3.2.3 provided an overview of many modeling algorithms and described each model's level of interpretability and explainability. However, each of these algorithms also have a perceived level of accuracy of performance. Barredo Arrieta et al. (2020) already described the perceived model accuracy, and this chapter presents mappings on the perceived accuracies to their respective modeling algorithms in Fig. 3.7b. While there are some differences between the presented mappings in this chapter and Barredo Arrieta et al. (2020)'s on a small number of specific types of models, one larger shared relationship is present: models that are less interpretable have higher accuracy and models that are more interpretable are less accurate.

For example, both mappings agree that OLS is perceived to be more inaccurate than CNNs and that CNNs are less interpretable than OLS. However, it would be more precise to state that these are perceived accuracies of these models and not verified truths.

This chapter lays out the argument that these perceived accuracies are, in fact, severely flawed. It is inappropriate to cast these broad overarching statements on the capacity of a particular modeling algorithm to accurately capture the true underlying relationship without contextualizing the problem or considering the true underlying relationship of the phenomena of interest. For example, in the scatterplot with the linear regression line of best fit of the first dataset in Fig. 3.7a has data that bounces along the estimated model. However, with the amount of data presented, it is unclear if the true relationship is linear with some noise or actually follows a cyclic or polynomial-like pattern. Thus stating that OLS or NLM is more accurate than the other for this example is inappropriate. The modeler must assume or have some subject matter expertise to make a proper judgment about which modeling technique is more accurate. Further, Section 3.3 will show examples where XAI methods are in fact able to outperform CNNs, which provide two counterexamples to the perceived model accuracies of the modeling community.

3.2.4 Dimensionality reduction

Despite one's best efforts, one might not know which variables are important for a given task. Even if all of the variables are interpretable and explainable, it may be difficult to describe how these variables interact with one another in a meaningful manner. Thus final models can be challenging to unravel when the number of variables is too large. For example, this occurs in economics during expert failure, where experts model a complex phenomena using features (Murphy et al., 2021). Thus dimensionality reduction techniques attempt to simplify the amount of features needed to describe the data. Simpler methods reduce the number of variables used, whereas others find a latent space for describing the data. If similar evaluation metrics and model predictions are obtained using the latent space

or the model with less variables when compared to the model with all the variables in the original space, then the simpler model is often preferable.

3.2.4.1 Subset selection procedures

Many of the first attempts at reducing the number of variables used in a model use rules to reduce the number of variables in a model. The simplest of these rules involved using the top most important variables from a model with all of the variables, and then rebuilding the model with only those variables. A variation of this approach could try all possible combinations of a specified number of features. This procedure is called the best subset selection (Draper and Smith, 1998; James et al., 2013). This approach is generalizable to many different modeling algorithms.

More sophisticated approaches than using the most important variables involve iterative algorithms, which add or remove variables until a specified value converges or other algorithmic criteria are satisfied (Draper and Smith, 1998; James et al., 2013; Mendenhall and Sincich, 2011). Many of these approaches were developed for linear models, so the extent of their applicability is somewhat limited. Some of these approaches are the backward elimination and stepwise selection procedures (Draper and Smith, 1998). Based on the framework from Draper and Smith (1998), the framework for the backward elimination procedure is as follows:

1. Select an evaluation metric threshold value, α_0.
2. Build a model with all of the m variables.
3. Build a model with all of the variables except one. Repeat this for every variable to obtain m models.
4. Select the model from (3.) with the smallest evaluation criterion, α_1:
   - If α_1 satisfies α_0, then use the simpler model.
   - If α_1 does not satisfy α_0, then replace the original model with the model that produced α_1. Repeat step (3.).

Draper and Smith described this procedure using the F statistic from a regression model. However, this procedure is generalizable to many different modeling algorithms, such as SVM. The basic idea is to build an initial model with all of the variables and calculate a metric for evaluation with

a set criterion. The algorithm then continually removes variables until the criterion is satisfied.

There are many similar methods to the stepwise procedure, such as backward elimination. In general, these subset selection procedures are interpretable and explainable since every step in the algorithm can be interpreted and explained.

3.2.4.2 LASSO, ridge, and elastic net

A popular method in the statistical learning community is the use of the least absolute shrinkage and selection operator (LASSO) on generalized linear models (Hastie et al., 2017; James et al., 2013). A similar method is the Ridge (Draper and Smith, 1998; Hastie et al., 2017; James et al., 2013). However, both are equivalent to the elastic net (Hastie et al., 2017). These methods shrink the variables of the model to 0, and retains only the important ones.

For the LASSO, for a given objective function, the LASSO is the solution

$$min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} f(x_i, y_i, \alpha, \beta), \tag{3.51}$$

subject to $||\beta||_1 \leq t$, where $||\cdot||_1$ is the L_1-norm, and t is a tuning parameter (Tibshirani, 1994). Ridge provides a similar solution, except it is subject to $||\beta||_2^2 \leq t$, where $||\cdot||_2$ is the L_2-norm (Hastie et al., 2017). The elastic net uses the same object function as the LASSO and Ridge, but it is subject to $(1-\delta)||\beta||_1 + \delta||\beta||_2^2 \leq t$ such that $\delta \in [0, 1]$. Thus the LASSO and Ridge are a special case of the elastic net (Hastie et al., 2017).

One way to understand the LASSO and Ridge is through the use of Bayesian updating (Tibshirani, 1994). The Ridge assumes that the variables have a normal prior, whereas the LASSO assumes that the prior is a Laplace. Similarly, the elastic net has a prior that compromises between the Gaussian and Laplace (Zou and Hastie, 2005). Thus one is able to have a highly interpretable and explainable variable selection algorithm since one is merely performing Bayesian updating.

In practice, one would typically build a model using the elastic net and obtain a subset of variables. Once the analyst was satisfied with the subset

of variables, the analyst would rebuild the model using only that subset of variables. Then the model has the typical interpretation and explanation. Thus this family of methods are very powerful and has been well developed. However, it is not applicable to all modeling algorithms.

3.2.4.3 PCA

Principal component analysis (PCA) is a fundamental technique in a variety of fields. Furthermore, we discussed much of the technical details of eigenvalues for eccentricity. However, PCA also uses eigenvectors to provide a latent space of the original data. Summarizing the descriptions from Lattin et al. (2003) and Izenman (2008), PCA attempts to model describe

$$\epsilon_j = \sum_{i=1}^{r} b_{ji} X_i, \forall j \in \{1, ..., t\}, \tag{3.52}$$

such that X_i is the i^{th} vector where $i \in \{1, ..., r\}$, $t \leq r$, ϵ_j is the j^{th} principal component or score, and b_j is the j^{th} the eigenvector (Hotelling, 1933). The theoretical underpinnings of PCA allow us to state that the eigenvectors of the given data, λ_j, explain the variation in the data (Hotelling, 1933). For example, one can state that the first principal component accounts for $\frac{\lambda_1}{\sum_{j=1}^{r} \lambda_j} \times 100\%$ of the variation in the data. Extended details on PCA's theoretical foundations and computational methods can be found in Hotelling (1933) and Izenman (2008). The resulting latent space is composed of the ϵ_j's, where analysts usually pick 2 (which makes $t = 2$) so that a 2D scatterplot can be constructed. More formally, the number of components to retain should be determined using an elbow or scree plot or Kaiser's rule (Kaiser, 1960). A scree plot includes the numeric value the eigenvalues have on the y-axis and the order of each of the eigenvalues. Scree plots are helpful when there is an obvious large deviation in the amount of variation explained by the components. Kairer's rule is less human dependant, and stipulates that all of those principal components with eigenvalues greater than or equal to 1 should be retained (Kaiser, 1960). However, other guidelines state that the amount of variation explained by the retained principal components should be at least 50%. Thus there are many tools for assessing

the results of PCA, but there is not a "best" method. Thus careful considerations must be made when using PCA. PCA is explainable since it is the linear combination of the data that captures the data. PCA is somewhat interpretable since one is able to assign meaning to each latent space by interpreting the loadings.

This approach preserves the variation in the data, so it is able to describe it in a linear manner. Therefore PCA is able to construct a smaller feature space to describe the data. It is important to note that PCA often uses the correlation matrix of the data to capture these eigenvalues and eigenvectors (Lattin et al., 2003). Thus much of what is discovered using PCA is exploratory and not definitive.

One is able to analyze the eigenvectors to describe what that dimension of the space means (Lattin et al., 2003). For instance, it may be capturing all of the shape metrics in the first dimension of the PCA space and the color metrics in the second. However, there isn't a defined mathematical method for analyzing the eigenvectors. It takes practice and intuition to learn how to describe these eigenvectors. Therefore while one is able to interpret and explain PCA, the interpretations of PCA are not precise and straightforward.

3.2.4.4 FA

Factor analysis (FA) is similar to PCA (Izenman, 2008; Lattin et al., 2003). Sometimes to estimate FA, PCA is used instead (Lattin et al., 2003). Thus the discussion on FA will be brief. However, the theoretical differences essentially amount to a rotation component (Izenman, 2008; Lattin et al., 2003). When this subtle difference is included in the estimation of FA, it can help to further separate different groups from one another. For example, it may make it more obvious when interpreting the loadings that the first loading corresponds to shape metrics and the second to color. However, this approach is controversial, as it may not generalize well (Blackith, 1971; Manly, 1994).

The inclusion of the rotation parameter matrix makes FA less explainable since the choice of estimating the rotation is not straightforward. Though there are popular choices, such as the varimax rotation (Izenman, 2008; Kaiser, 1958; Lattin et al., 2003), the choice of the rotation matrix drastically

changes an analyst's ability to explain the results. The interpretability of FA is essentially the same as the reasons for PCA.

3.2.4.5 Fourier transform

The Fourier transform (FT) is a well studied technique from signal and image analysis (Gonzalez et al., 2009; Kinser, 2018; Russ, 1995). Using complex numbers, the FT is

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{-iux} dx, \tag{3.53}$$

where u represents the frequency space values and $i = \sqrt{-1}$. It is important to note that

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{3.54}$$

Thus FT is essentially decomposing a signal into the summation of cosine and sine waves. One is able to interpret the magnitude of $F(u)$ as the magnitude of u that is in $f(x)$ (Kinser, 2018). It is also well known that

$$f'(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(u) e^{iux} du. \tag{3.55}$$

Note that $f'(x) = f(x)$. Thus no information is lost when projecting into the Fourier space (Kinser, 2018). These relationships can be extended to the case where one has two variables (such as one does with images). For this one has that

$$F(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-i(ux+vy)} dx dy, \tag{3.56}$$

$$f'(x, y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(u, v) e^{i(ux+vy)} du dv. \tag{3.57}$$

The FT is extremely powerful, and is used for a variety of image processing tasks, such as noise reduction, dimensionality reduction, and data compression (Gonzalez et al., 2009; Kinser, 2018; Russ, 1995). Though these approaches are not extensively presented in traditional machine learning or statistic textbooks, a brief introduction to this technique is valuable for those readers analyzing image data.

Much work has been done to interpret and explain FT. However, since it is primarily limited to signal processing, the discussion will be restrained. FT is somewhat explainable since one is able to describe it's process precisely, but it uses complex numbers. Simply stating that complex numbers have the property that $i^2 = -1$ does not make complex numbers explainable. The key component which needs to be explicitly described is $\sqrt{-1}$. Furthermore, since $\sqrt{-1}$ does not have any physical meaning, FT is only somewhat interpretable. Since many who use models may not have seen the FT or even complex numbers before, many audiences may struggle to understand this approach. Furthermore, FT assumes that the signal analyzed is repeating just like a cosine or sine function. This may not be an accurate representation of the data. Even with this caveat, the FT is still a useful technique in a variety of situations. For an extensive description of FT, we recommend reading Kinser (2018); he presents many examples alongside the theoretical descriptions of image processing methods in Python.

3.2.4.6 Manifolds

There are a number of techniques based on estimating manifolds for dimensionality reduction, such as uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) and t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). These techniques have recently gained popularity in a variety of communities, such bioinformatics and genomics (Alexandrov, 2020; Bravo González-Blas et al., 2020; Takei et al., 2021).

Part of the issue with these methods is that a straightforward definition of a manifold is difficult at best to describe. Izenman (2008) stated that "It is not easy to give a simple description of a 'manifold' because of the complex mathematical notions involved in its definition." He attempts to give a simple definition of a manifold as "a topological space that locally looks flat and featureless and behaves like Euclidean space." However, two prime examples of manifolds are S-curves and swiss rolls (Izenman, 2008). These objects are not flat in 3D space, yet many traditional clustering or dimensionality reduction techniques struggle to represent these objects well. Thus while one can present examples of manifolds and describe them with

mathematical notation, one still has difficulty in describing them. Therefore trying to estimate something that cannot be explained or interpreted easily is setting up the procedure for estimating the object much more difficult to explain and interpret as well.

Furthermore, UMAP, t-SNE, and similar manifold-based approaches for dimensionality reduction do not preserve global distances nor do they loadings by which to interpret or explain their results (McInnes et al., 2020). In addition, it is unclear whether these methods will perform well in small data scenarios. It should be noted that small data scenarios are not from a bygone era, but are still a part of the age of "big data" (Lamberti, 2020b). For example, we may have millions of observations in a given dataset, but only a handful (less than 10) of a particular class. These rarer cases may be of extreme importance, such as terrorist attacks or a rare case for an illness. One would not want to casually label such cases as an outlier and remove them from the data. Thus one would also not want to use a technique to describe cases that one has very few observations. This does not mean that approaches, such as UMAP, should never be used. Data may be so untenable that using an uninterpretable and unexplainable solution may be the best an analyst is able to provide. However, using these approaches comes with a cost, and an analyst should be aware of the benefits and weaknesses of these approaches. In essence, manifold-based approaches for dimensionality reduction are unexplainable and uninterpretable since one cannot clearly describe a manifold and ascribe physical meaning to their latent spaces, respectively.

3.2.5 Model assurance

There are a number of ways to provide model assurance, such as resampling methods, effect comparison, analysis of influential observations, human-in-the-loop (HILT) models, and visualization methods. We want to build models that are generalizable so that the models are useful beyond the data that they were built with high levels of confidence and trust. Model assurance confidence is increased by using explainable and interpretable models and metrics. Using model assurance methods alongside interpretable and explainable features and modeling algorithms provides the analyst insight

into the model's predictions and internal logic. Using model assurance also helps to identify a clear path to improve the model or to collect different features.

Model assurance differs from data assurance. We define **data assurance** as processes which help to evaluate that the sample data accurately describes the population data. Similar to increasing confidence for model assurance methods, adhering to interpretability and explainability increases confidence in the sample data. Thus the primary difference is the subject: the data for data assurance and the model for model assurance. Model and data assurance are the two main pillars of AI assurance. Interested readers in AI assurance are invited to read Batarseh et al. (2021) for extended details and discussions.

3.2.5.1 Resampling methods

Resampling methods are used in a variety of fields to verify that a model is robust. Though collecting more data would be the most obvious way to check if a model is robust, this cannot always be possible. Various costs, such as time or money, may make collecting data untenable. Thus model validation approaches are developed to verify that models are robust with limited amounts of data. There are three primary model validation methods: training-validation splits (James et al., 2013), cross validation (CV) (James et al., 2013; Kohavi, 1995; Takei et al., 2021), and the bootstrap (James et al., 2013; Kohavi, 1995; Takei et al., 2021).

One need to define some terminology before one discusses the specifics of CV. Be aware that different communities define these terms differently and may interchange their definitions. Thus it is vitally important to ensure that these terms are defined explicitly. **Training data** is the data used to build the model. **Testing data** is data used to check the model and may be used to train the model. **Validation data** is data used to check the model, but never used to train the model. Before any model is built, the data is usually randomly allocated to the training and validation data using a 70% and 30% split from the original data (James et al., 2013). However, it is not uncommon to observed different values from these splits (Lamberti, 2020a; Li et al., 2019). This method is usually called the holdout method (James et al., 2013).

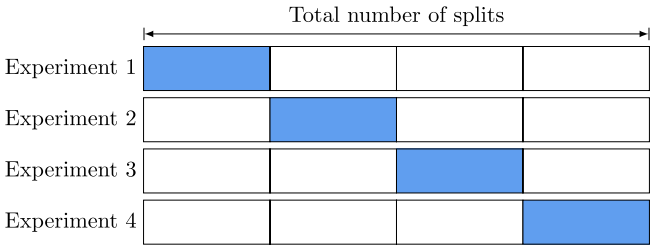


FIGURE 3.6 Figure is a 4-folds CV visualization.

The next approach is called k-folds CV, where the training data is split into k approximately equal parts, which take turns being the training and testing data (Kohavi, 1995). This is exemplified in Fig. 3.6. For example, in the first experiment, the data is built using all of the data except the first split (the blue (dark gray in print version) partition). The first split is then used to check the quality of the model. For the second experiment, all of the data is used except for the second split. The second split is then used to check the quality of the model. This process is continued for each split. A good model would have similar values across all of the splits.

Sometimes k-folds CV is combined with the holdout method. This combination first creates the training and validation data. k-folds CV is then applied to the training data. A model is built using all of the training data, and then checked using the validation data. A robust model would have similar model evaluation metrics cross all of the folds and for the validation data.

The bootstrap works by treating the collected data as an empirical estimate of the population's distribution (Efron, 1979; James et al., 2013). The observed data is randomly sampled with replacement to a specified sample size, n. A model is built using this estimate and evaluated. This is repeated m times. The m evaluation metrics are then analyzed by ensuring that they have a small variance. A small variance ensures that the model is not changing and remaining consistent.

In computer vision, a popular approach to increase the number of observations in the dataset is to use data augmentation. This process is similar to the bootstrap since both are approaches used to increase the size of one's dataset. Data augmentation is essentially a series of transformations

applied to the images. The types of transformations vary, but include rotations, smoothing, flipping, and scaling (Mikolajczyk and Grochowski, 2018; Miller et al., 2000; Wang et al., 2017). Data augmentation is a useful technique for increasing the total number of observations in a given dataset, but the process of data augmentation is not applicable to non-image data.

3.2.5.2 Effect comparison

One of the primary ways to perform effect comparison is the use variable importance (VI) by comparing the VI values on two different datasets. For instance, calculating VI changes depending on the type of model used for the analysis. The VI for linear regression models is calculated using the absolute value of the coefficient's t statistic (Molnar, 2020), whereas a random forest model's VI is based on the Gini index (James et al., 2013). It is important to note that a model agnostic method does exist, but it is a post-hoc analysis (Breiman, 2001). Regardless, the final values for a given model with the same variables can always be compared with the same type of model. One can exploit this fact to measure the differences between different models. One would expect similar models to have very small differences between the VI measures, whereas different models would have large differences. Thus one would show evidence for model assurance when there are small differences between the VI values for separate models built on similar but different data. Using interpretable and explainable modeling algorithms and metrics are critical for this method to be useful. When the analysts cannot interpret and explain the algorithm's parameters and metrics, the effect comparison becomes less helpful for model assurance.

3.2.5.3 HILT models

Combining human and machine learning is another approach to ensure that a robust model is built (Lamberti, 2020b; Trzaskoma et al., 2020; Valen et al., 2016; Witten et al., 2011). This mixture of intelligences can help to bolster the other's weaknesses. Machine learning can provide the strict rules needed to generalize tedious and computationally heavy tasks. Human learning can guide computational methods to more accurate descriptions of a phenomena. However, these models can be obtuse to those who were not involved in the model building process. The human components are

also difficult for other humans to replicate unless the analysts provide clear and precise descriptions of the their choices. Therefore these types of models provide evidence in favor of deep learning methods (Caicedo et al., 2017; LeCun et al., 2015; Lundberg and Borner, 2019; Pärnamaa and Parts, 2017). Though HILT models provide useful initial steps for explaining a phenomena, replacing the human components with interpretable and explainable computational methods is highly encouraged.

3.2.5.4 Influential observations

Observations which impact model's parameters more than other observations are considered to be influential observations or outliers. Influential observations are typically considered important for a particular algorithm. For instance, SVMs define the support vectors as those observations that are needed to define the separating hyper-object (Hastie et al., 2017; James et al., 2013).

The term "outliers" usually has a negative connotation. It is not uncommon for these observations to be removed to provide more accurate estimates for parameters of interest (Mancl and DeRouen, 2001; Mendenhall and Sincich, 2011; Miller, 1974; Preisser and Qaqish, 1996). However, the removal of potential outliers is not straightforward (Draper and Smith, 1998). For example, this outlier could provide evidence that the initial data collection missed an entire subset of the population or indicates that a subpopulation is rare (Lamberti, 2020b; Wand et al., 2021; Wojcik et al., 2019). Conversely, if the goal of the analysis is to model the typical observations, then outliers should be removed. Though this might make the model provide more accurate predictions for those typical observations, rarer observations will have larger errors on this altered model. Thus great care must be taken when considering to remove outliers when the effects on the model's parameters change dramatically.

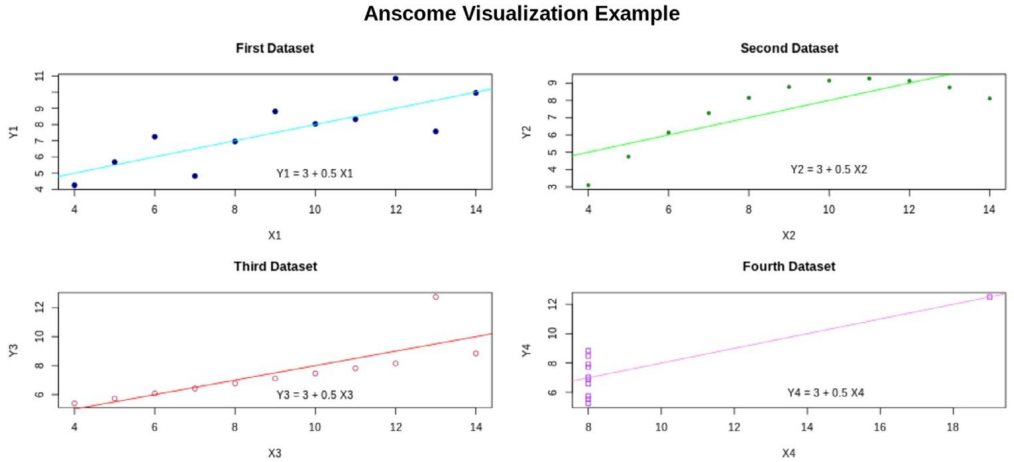
There are several methods for assessing influential observations. Data visualization techniques, such as boxplots can provide evidence of the presence of outliers (Peck et al., 2008). Leverage and Cook's distance are useful measures of an observation's influence on general linear and non-linear models (Myers, 2010; Preisser and Qaqish, 1996).

3.2.5.5 Visualization methods

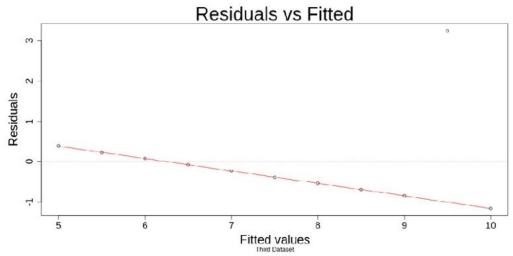
Using visualization tools can provide greater insight to an analyst's model and help to bolster the robustness of the model (Anscombe, 1973; Cleveland, 1993). Though this idea is not new, it still holds much relevance for model assurance. Furthermore, visualization methods are able to make variables or algorithms more explainable and interpretable. Both of these concepts are well illustrated using the famous Anscombe datasets (Anscombe, 1973).

The Anscombe datasets provide 4 pairs of explanatory and response variables. When the analyst builds a simple linear model for each of the four pairs, the resulting models are strikingly similar. All have similar values for the model's parameters, R^2 values of about 0.667, t-statistic values of about 4.24, and stand error estimates of about 0.118 (Anscombe, 1973). However, if one were to plot the data against the estimated model, one will observe concerning aspects, as seen in Fig. 3.7a. The first dataset is the only model that adequately captures the phenomena. The second data appear to follow a polynomial pattern, and thus a simple linear model is inappropriate. The third dataset does follow a linear pattern, except for one observation. It is unclear if that one observation is an outlier, error in the data collection process, or part of the phenomena of interest. The fourth dataset has one observation that is drastically changing the estimated model. Thus observing the statistics, model parameters, and other model evaluation metrics are usually not sufficient for determining if the model adequately captures the relationship in the data.

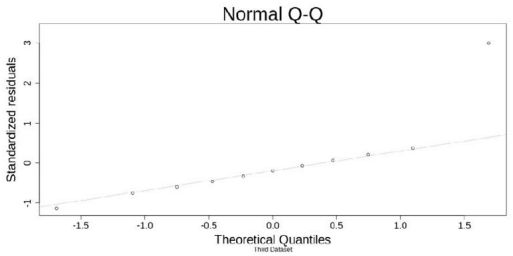
Though one cannot provide a complete overview of graphics for modeling and model assessment, using visualizations is an excellent method for ensuring the robustness of a model (Bhattacharyya and Johnson, 1977; Cleveland, 1993; Peck et al., 2008; Wickham, 2016). One of the fundamental methods for visualizing data is the scatterplot. By plotting the data, analysts might be able to quickly assess the relationship data follows. This assessment will help in selecting a modeling algorithm that would best describe the data. For example, one familiar with the art would select a non-linear (possibly parabolic) model for the second Anscombe dataset af-



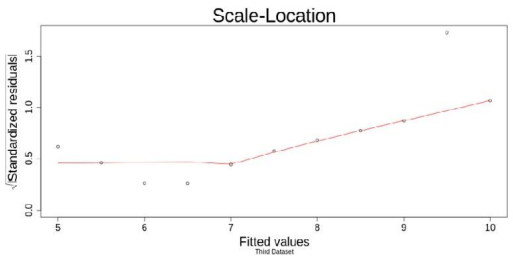
(a) The scatterplots and linear models for the famous Anscombe datasets



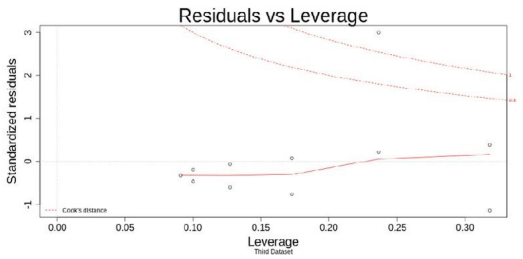
(b) Residuals versus fitted values plot for the third Anscombe data



(c) Normal Q-Q plot for the third Anscombe data



(d) Scale-location plot for the third Anscome data



(e) Leverage plot for third Anscombe data

FIGURE 3.7 Figure showcasing utility of visualizations for model assurance.

ter one observed the non-linear relationship displayed in the scatterplot in the topright of Fig. 3.7a with green (gray in print version) solid points.

Another useful tool based on scatterplots for assessing models is observing the relationship between the residuals and the predicted values of a

model (Draper and Smith, 1998). For OLS models, this plot should have a random scatter surrounding the residual axis at 0. Fig. 3.7b provides evidence that the OLS model built does not fit the data well. There are a variety of other visualization tools available for OLS, such as the scale-location plot in Fig. 3.7d and the residual versus leverage plot in Fig. 3.7e (Belsley, 2004; Mendenhall and Sincich, 2011).

If one is able to make assumptions about the parametric distribution of the data, one should use QQ-Plots to assess if the data does follow the assumed distribution (Peck et al., 2008). Since an OLS model was used for the Anscombe datasets, one is able to assess if the residuals follow a normal distribution using a normal QQ-plot. Data that follow a normal distribution will follow a straight line from the bottom left of the plot to the top right. Since Fig. 3.7c does not follow a straight line due to the observation in the top right of the plot above the reference line, one has strong evidence that the data does not follow a normal distribution.

Whisker-plots or boxplots are a useful visualization to assess the distribution of data (Cleveland, 1993; Peck et al., 2008; Tukey, 1977). Based on quantiles, they quickly provide a visual for analysts to obtain a high level view of the data. Potential outliers are indicated by singular points beyond the "whiskers" of the plot. They are those observations more extreme than the first or third quantile minus or plus, respectively, $1.5 \times IQR$, where $IQR = Q3 - Q1$ such that $Q1$ is the first quantile and $Q3$ is the third quantile. Fig. 3.10 in Section 3.3.2 provides an example of a boxplot. For an ideal symmetric distribution, one would expect both sides of the boxplot to be symmetrical with few, if any, potential outliers.

If an analyst is able to exploit an algorithm by using visualizations to describe how the model is making decisions, this makes the model much more explainable and interpretable. Visualizations that are able to describe all of the observations used to build and check the model are much more impactful than describing a visualization for a single observation. Furthermore, visualizations that use metrics that are universal across different models are much more interpretable and explainable than visualizations that describe a relative latent space. For example, describing how a model performs using the residual versus fitted values is much more interpretable

and explainable than a PCA plot of the first two principal components, since principal components must be interpreted for each unique dataset.

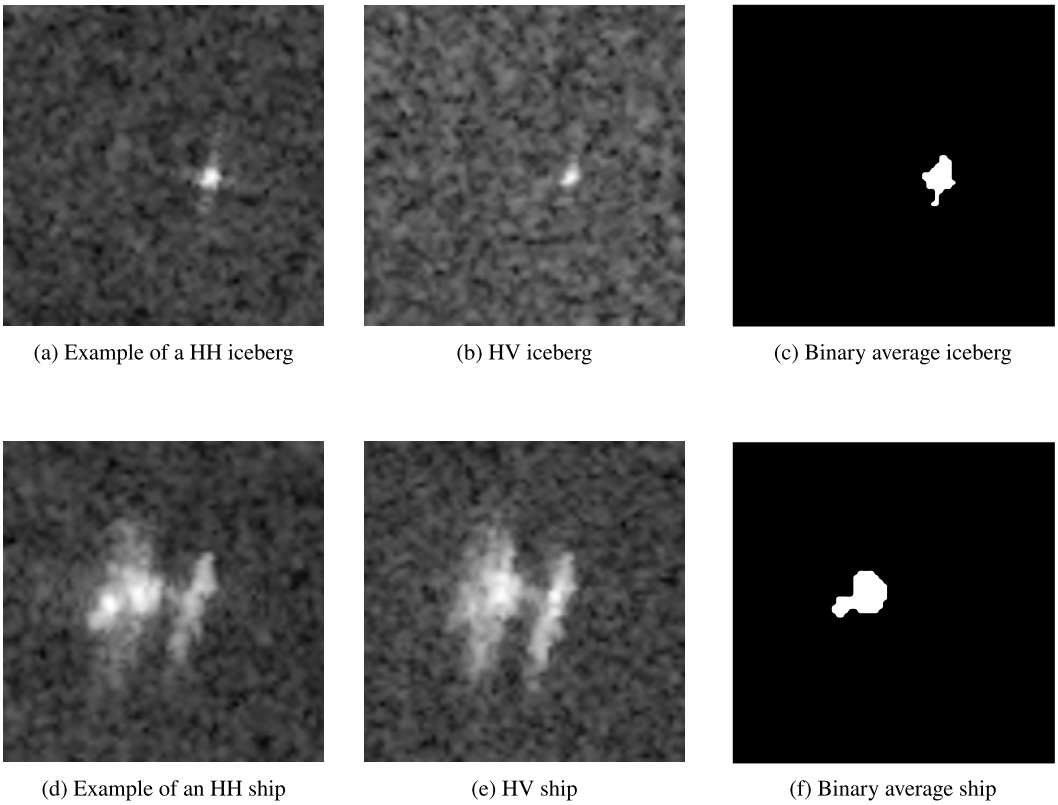
3.3 Experiments using XAI models

The following sub-sections provide examples of using XAI models. The first example is classifying satellite images of icebergs versus ships. The second example is classifying malignant versus healthy white blood cells (WBCs). The final models in both examples are compared to the state-of-the-art deep learning CNNs. Furthermore, the XAI models are able to outperform the deep learning methods. Thus XAI approaches are capable of tenable solutions when compared to deep learning solutions.

3.3.1 Satellite imagery

Lamberti (2020a) created a RF which provides highly accurate classification between icebergs and ships using satellite images. This corresponds to outperforming CNN-based approaches by about 7% and 11% on the testing and validation data, respectively. The first step is processing the images to extract the shapes of the objects. Shape metrics are then collected on the final objects. These shape metrics indicate that the important variables for discriminating these two classes are eccentricity and shape proportions (SPs). Readers interested in extended details should refer to Lamberti (2020a).

The data used in this model came from the "Statoil/C-CORE Iceberg Classifier Challenge" (Kaggle, 2017). These images are obtained from the Sentinel-1 satellite. The data was saved in `json` format, where each image was 75×75 pixels with two bands. The bands are radar backscatter produced from different polarizations. The first band corresponds to transmitting and receiving the waves horizontally (HH). The second band corresponds to transmitting the waves horizontally and receiving the waves vertically (HV). Each pixel value corresponds to real numbers in dB units. Examples of these images and the resulting extracted objects are provided in Fig. 3.8. There are 851 ships and 753 icebergs. This corresponds to a total of 1604pairs of images.



(a) Example of a HH iceberg (b) HV iceberg (c) Binary average iceberg

(d) Example of an HH ship (e) HV ship (f) Binary average ship

FIGURE 3.8 Figure provides examples satellite images and the final extracted shapes.

After performing 10-fold CV, the tuned parameter for the number of variables to check was 6. This resulted in an accuracy of about 95%. We then used the complete testing data to build a RF model with the tuned parameter value. This resulted in an accuracy of just over 99% and 95% on the testing and validation data, respectively. This is summarized in Table 3.2. The CV error presents a much more accurate estimate of the validation data accuracy rate as they only differ by about 1%.

The confusion tables for the testing and validation data are provided in Table 3.3. The model classified more icebergs incorrectly on the testing data and misclassified more ships on the testing data. Nonetheless, the model was able to correctly classify both classes with a high level of accuracy.

Table 3.2 The table depicts accuracy of the CV accuracy estimate, the complete testing data, and the validation data. Notice that the CV estimate is very similar to the testing data estimate.

Data	Accuracy
CV	95%
Testing	>100%
Validation	96%

Table 3.3 Table is the confusion matrix for the final model using all of the testing data. The validation data is in parentheses and boldened. This corresponds to an overall accuracy just under 100% and about 96% on the training and validation data, respectively.

Prediction/Reference	Ship	Iceberg
Ship	2721 **(650)**	18 **(25)**
Iceberg	3 **(30)**	2392 **(577)**

Table 3.4 Table provides the rescaled importance of each of the variables; 100 means that the variable is the most important and 0 means that the variable is the least important. The 2 most important variables were eccentricity and SP.

Metric	Variable Importance
Eccentricity	100.00
SP	86.56
EI: White	85.84
EI: Black	83.35
Circularity	79.01
Rectangularity	65.27
White Bounding Box	49.45
Black Bounding Box	42.28
1^{st} Eigenvalue	0.60
2^{nd} Eigenvalue	0.00

The relative importance of the variables are provided in Table 3.4. The importance was calculated using the decrease in the mean number of correctly classified observations when that given variable was removed from

Table 3.5 Table includes the CNN results from Li, Huang, Peters, and Power and the RF from Lamberti on the testing and validation data (Lamberti, 2020a; Li et al., 2019). The RF model outperforms all of the CNN-based approaches.

Metric (%)	CNN 1	CNN 2	CNN 3	CNN 3	RF	Data
Accuracy	91.5	92	93.5	94.9	>100.0	Train
MP	92	94	94	94	>100.0	Train
Mean Recall	92	94	94	94	>100.0	Train
Mean F_1-Score	92	94	94	94	>100.0	Train
Accuracy	86.51	87.72	84.83	87.02	95.71	Valid
MP	87	88	85	87	95.68	Valid
Mean Recall	87	88	85	87	95.72	Valid
Mean F_1-Score	86	88	85	87	95.70	Valid

the model. The 2 most important variables in the model were eccentricity and SP.

We provide the summary of the CNN and RF models' accuracy, mean precision (MP), mean recall, and mean F_1-score for the testing and validation data in Table 3.5. This corresponds to a 7% and 11% mean outperformance when using overall accuracy for the RF model.

3.3.2 White blood cell

Lamberti (2022) created a RF model, which provides highly accurate classification between malignant and healthy white blood cells (WBCs). State-of-the-art solutions use a CNN-based solution using a large number of features (Sahlol et al., 2020). However, Lamberti was able to outperform these solutions with dramatically less features, while using a more explainable and interpretable XAI algorithm. Since one is able to extract the VI, he was able to show that the batch effects have a strong effect on the model. These effects must be considered when building a model for deployment at a clinical level. Readers who want additional details should refer to Lamberti (2022).

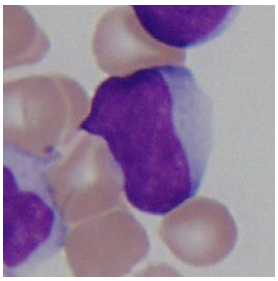
The two sources of data used in this example were the publicly available ALL-IDB (Labati et al., 2011) and C-NMC (Duggal et al., 2016; Gupta et al., 2017) datasets. The ALL-IDB data was provided by the Department of Information Technology - Università degli Studi di Milano. The image data

was captured with a microscope with a Cannon PowerShot G5 camera and are retained in JPG format with 24 bit color depth (Labati et al., 2011). The files were received as TIFs. The ALL-IDB source has two datasets, but only the ALL-IBD2 data will be utilized. The ALL-IDB2 data contain cropped areas of interest of WBCs that are malignant and healthy. These cells retain the background and other potential nearby cells. There are a total of 130 of malignant and healthy cells, for a total of 260 images. Examples of the ALL-IDB2 dataset are provided in Figs. 3.9a and 3.9c.

The C-NMC dataset comes from the ISBI 2019 challenge. The images are saved in the BMP format. There are a total of 7272 malignant cells and 3389 healthy cells for a total of 10,661 images. The cells were extracted by an expert oncologist. Examples of the C-NMC dataset are provided in Figs. 3.9e and 3.9g. By using two datasets, one is able to compare the VI for the two separate RF models. When one compares each model's VIs, one can characterize the batch effects between the two sources.

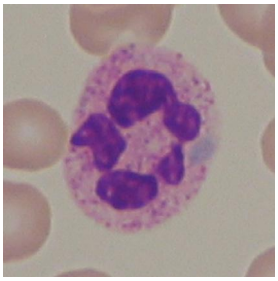
An important distinction between the ALL-IDB2 and C-NMC datasets is that they both have differing backgrounds. This is crucially important as the segmenting algorithm Lamberti developed can be applied to both datasets without any changes. This provides evidence that the segmentation algorithm is generalizable to other data sets. Examples of the resulting segmentation results are provided in Fig. 3.9.

A variety of non-shape metrics related to the color and texture of the extracted objects for classifying the WBCs were selected. The mean and standard deviation of the amount of red, green, blue, cyan, magenta, yellow, and black in the objects were chosen as metrics to capture various aspects of the color of the cells. This corresponds to a total of 14 color metrics. The mean and standard deviation of the co-occurrence matrix was used. The co-occurrence matrix captures how often grayscale intensity values are next to other intensity values (Kinser, 2018). Large values indicate that a given grayscale intensity is usually next to another given grayscale intensity, whereas small value means that it is rarely next to another given grayscale intensity. Therefore a large mean indicates that values tend to be rough as intensities are nearby many different values. A small mean would indicate



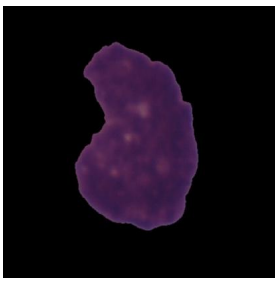
(a) Example of a malignant WBC from the ALL-IDB2 data.

(b) Binary shape extracted from the malignant ALL-IDB2 example in Figure 3.9a



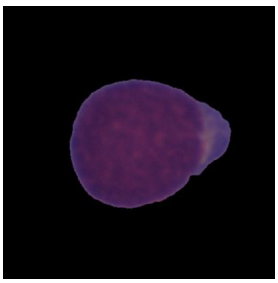
(c) Example of a healthy WBC from the ALL-IDB2 data

(d) Binary shape extracted from the healthy WBC example in Figure 3.9c.



(e) Example of a malignant WBC from the C-NMC data

(f) Binary shape extracted from the malignant C-NMC example in Figure 3.9e



(g) Example of a healthy WBC from the C-NMC data

(h) Binary shape extracted from the healthy WBC example in Figure 3.9g

FIGURE 3.9 Figure provides examples from the ALL-IDB2 and C-NMC datasets with their corresponding extracted shapes.

Table 3.6 Comparison with related works for the WBC datasets. The best approaches for the smallest number of features and highest accuracy are boldened. Note that # = number, Exp. = Explainability, Inter. = Interpretability, and Acc. = Accuracy.

Data	Source	# of Features	Model	Exp.	Inter.	Type	Acc.
ALL-IDB2	Singhal and Singh (2014)	256	SVM	Low	Low	AI	89.72%
	Singhal and Singh (2016)	4096	Knn	Low	Low	AI	93.84%
	Bhattacharjee and Saini (2015)	**8**	Knn	Low	Low	AI	95.24%
	Sahlol et al. (2019)	45	Knn	Low	Low	AI	95.67%
	Sahlol et al. (2020)	1087	CNN& SVM	Low	Low	AI	96.11%
	Lamberti (2022)	24	RF	High	Medium	XAI	**100.00%**
C-NMC	Kulhalli et al. (2019)	25×10^6	CNN	Low	Low	AI	85.7%
	Ding et al. (2019)	87×10^6	CNN	Low	Low	AI	86.7%
	Marzahl et al. (2019)	11×10^6	CNN	Low	Low	AI	86.9%
	Sahlol et al. (2020)	1115	CNN&SVM	Low	Low	AI	87.9%
	Lamberti (2022)	**24**	RF	High	Medium	XAI	**90.1%**

Table 3.7 The relative VI of the categories for the ALL-IDB2 and C-NMC data.

Relative Importance	Most Important	Secondarily Important	Least Important
ALL-IDB2	Color: 1.000	Shape: 0.357	Texture: 0.052
C-NMC	Shape: 1.000	Color: 0.776	Texture: 0.152

that intensities tend to be smoother. The standard deviation indicates how much variation there is in the typical grayscale intensity value.

Table 3.6 provide the results for the presented approach compared to other state-of-the-art approaches for the ALL-IDB2 and C-NMC datasets. Both of these models used the same preprocessing steps to ensure a fair treatment of the data. The presented model outperformed the other state-of-the-art approaches by about 6.31% and 3.81%, on average, for the ALL-IDB2 and C-NMC datasets, respectively.

Table 3.7 provides the relative variable importance (VI) by category. The most important category is color; shape is secondary, and texture is the least important. Table 3.8 provides the two relative VI for the two datasets. Table 3.9 compares the VI for the two datasets. Tables 3.8 and 3.9 show that the VI changes between the two datasets. This provides strong evidence in

Table 3.8 The table provides the relative VI of the features for the ALL-IDB2 data in the first two columns and the C-NMC data in the second two columns. The most important feature for the ALL-IDB2 was the mean pixel value in the blue channel, followed closely by the mean pixel value in the black channel, and then circularity. The most important feature for the C-NMC data was the White EI value, followed closely by the Black EI and Eccentricity.

Feature	Relative VI	Feature	Relative VI
Mean Magenta	1.000	White EI	1.000
SD Black	0.791	2^{nd} Eigenvalue	0.664
SD Blue	0.660	1^{st} Eigenvalue	0.614
2^{nd} Eigenvalue	0.539	Mean Blue	0.436
White EI	0.458	Black EI	0.427
Mean Green	0.393	Mean Black	0.414
SD Magenta	0.253	Mean Red	0.265
Mean Black	0.229	Mean Co-Occurrence Matrix	0.250
Mean Blue	0.214	Mean Green	0.248
SD Green	0.209	SD Magenta	0.247
Circularity	0.207	SD Co-Occurrence Matrix	0.245
SD Cyan	195	Number of Corners	0.204
1^{st} Eigenvalue	0.146	SD Black	0.176
SD Red	0.139	Mean Magenta	0.175
Mean Co-Occurrence Matrix	0.124	SD Blue	0.165
Number of Corners	0.105	Circularity	0.162
SD Co-Occurrence Matrix	0.097	SD Green	0.154
Mean Red	0.087	SP	0.129
SD Cyan	0.066	Mean Cyan	0.103
SP	0.057	SD Red	0.088
Black EI	0.052	Eccentricity	0.060
Eccentricity	0.023	SD Cyan	0.056
Mean Yellow	0.001	Mean Yellow	>0.000
SD Yellow	>0.000	SD Yellow	>0.000

the batch effects between these two datasets. Furthermore, the mean and standard deviation of the absolute value of the difference of the VIs between the two datasets are about 0.017 and 0.019, respectively. The boxplot of these values are shown in Fig. 3.10. This showcases that there is a mea-

Table 3.9 The VI of the features for the ALL-IDB2 and C-NMC data.

Feature	Mean Decrease in Accuracy for ALL-IDB	Mean Decrease in Accuracy for C-NMC
White EI	0.039	0.083
Black EI	0.004	0.035
SP	0.005	0.011
Circularity	0.018	0.013
Eccentricity	0.002	0.005
1^{st} Eigenvalue	0.012	0.051
2^{nd} Eigenvalue	0.046	0.055
Number of Corners	0.009	0.017
Mean Red	0.007	0.022
SD Red	0.012	0.007
Mean Green	0.033	0.021
SD Green	0.018	0.013
Mean Blue	0.018	0.036
SD Blue	0.056	0.014
Mean Cyan	0.017	0.009
SD Cyan	0.006	0.005
Mean Magenta	0.085	0.015
SD Magenta	0.021	0.021
Mean Yellow	>0.000	>0.000
SD Yellow	>0.000	>0.000
Mean Black	0.019	0.034
SD Black	0.067	0.015
Mean Co-occurrence	0.010	0.021
SD Co-occurrence	0.008	0.020

surable difference between the features in the two datasets. Thus analysts must consider batch effects when using models in deployment.

3.4 Discussion

XAI is preferred over AI approaches in many critical, significant, and life-altering applications since XAI methods have clearer explanations and interpretations. Though it may be tempting to believe the results AI meth-

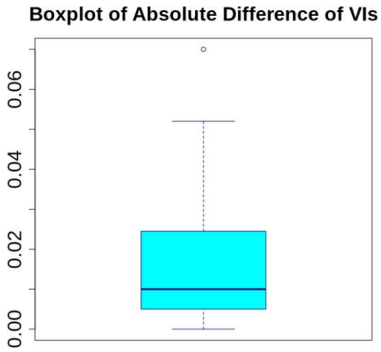


FIGURE 3.10 Figure provides boxplot of the absolute value of the difference of the paired VIs between the two WBC datasets.

ods in all applications, XAI incorporates interpretability, explainability, and model assurance to produce models that have fundamentally the same amount of utility as AI approaches. For instance, the examples presented in the previous section showcase various instances of XAI models employing interpretability, explainability, and model assurance in various capacities. Furthermore, both examples showcased XAI models that were able to outperform AI-based approaches. Although both XAI and traditional AI methods have human components, they differ in how the human interacts with the specific algorithm. In addition, though each example analyzed image data, the concepts presented can be easily extended to non-imagining problems.

3.4.1 XAI vs. AI in critical applications

Critical applications need to be explainable and interpretable. Having these qualities is not only necessary for the right to explainability, but are also vital for moral and ethical reasons. Obtaining a prediction which could have major impacts on an individual's life or livelihood need to be carefully explained and have clear precise interpretations, because getting these predictions wrong has dramatic repercussions in critical applications. While AI methods are able to provide evidence that a more interpretable and explainable solution exist, they do not provide clear insights as to what influences individual predictions. Thus XAI solutions are preferred for critical applications.

3.4.2 Explainability, interpretability, and model assurance in practice

A RF model was used in both applications and employed explainable and interpretable features and methods alongside model assurance techniques. For example, both approaches utilized the model assurance technique of stratification during the training-validation split. The RF model has high explainability and medium interpretability. Since a RF model was used, one could extract the VI. Furthermore, since interpretable and explainable metrics were used, one can state what aspects of the objects were important for classifying the different groups from one another. We also utilized a boxplot to showcase the difference between the WBC data VI values. This provided evidence that the data differs from each another.

3.4.3 XAI models outperform CNN-based solutions

The XAI RF model for satellite images outperforms the CNN-based approaches by about 7% and 11% on the testing and validation data, respectively, when we used the overall accuracy as the metric of choice. This showcases the power of using XAI features, which accurately describe the icebergs and ships.

The XAI RF model for classifying malignant and healthy WBCs was able to outperform the other advanced methods by about 5.20% on average. Furthermore, the presented approach used the smallest number of features when compared to the other approaches except for one. This is particularly impressive since the methods to extract the features were the same on both WBC datasets. Thus the presented approach is applicable to a large variety of different kinds of WBC data sources.

Therefore stating that deep learning methods are superior since they are able to outperform XAI is inaccurate. As one has shown in the examples, XAI models are able to outperform the state-of-the-art deep learning models.

3.4.4 XAI, deep learning models, and human inputs

XAI and deep learning models both require human analysts. The deep learning models need to be impregnated with an architecture and an appropriate method for learning features. XAI requires the human to select

relevant features, an appropriate modeling algorithm, and/or hyperparameters. Both XAI and deep learning require human analysts, but require different actions from those analysts.

3.4.5 Extending the lessons learned to non-image problems

Much of what we discussed in this chapter is directly applicable to non-image-based problems. The metrics extracted from the images were all scalar values that are used in XAI models. However, the concepts of explainability and interpretability we applied to the presented shape metrics can be directly applied to other metrics. For instance, one could examine gross domestic product (GDP) in terms of its explainability and interpretability.

3.5 Future work

Incorporating deep learning methods using a DAMG implementation may provide greater insight to the deep learning decision process for classification problems. This could help deep learning methods by simplifying the classification tasks. This could in turn reduce the number of needed features to learn, as there would only be a binary classification task. This could then allow for a simpler deep learning architecture. Visualizing the layers of these deep learning models could help with interpreting the results. Furthermore, each child deep learning model could use transfer learning from the parent deep learning model to learn the needed features quicker. This could help to make deep learning solutions more interpretable and explainable.

Though interpretability and explainability are important topics, it is difficult to quantify them. This is similar to assigning a numeric value to "joy," "pain," or "intelligence." There are systems designed to quantify some of these concepts, such as a grade point average (GPA), but it is difficult to quantify interpretability and explainability. Additionally, systems such as GPAs are not without their issues. One could argue that GPAs merely capture an individual's ability to regurgitate what teacher expects a student to know in a classroom setting. In other words, it does not directly capture a student's intelligence. Thus though a student's GPA and intelligence may correlate positively with one another, GPA is not a true measurement of

intelligence. Nevertheless, the design of a metric or index that would encapsulate interpretability and explainability would provide a useful guideline to quantify these concepts.

3.6 Conclusion

Explainability and interpretability are key components of XAI. They provide a foundation for model assurance methods, such as effect comparison and influential observations. Furthermore, the provided examples showed that the XAI models are able to outperform deep learning approaches using interpretable and explainable metrics for image data in health sciences and satellite analysis. These examples show that the perceived accuracies of various modeling approaches do not hold. The concepts discussed easily extend to non-imaging-based problems, which extends the impact of explainability and interpretability. Thus explainability and interpretability are powerful concepts to help analysts provide robust and generalizable models.

Acknowledgments

We would like to acknowledge Josef Lamberti for his writing tutelage. Jon Murphy and UVA Engineering Graduate Writing Lab Peer Review Group also provided valuable feedback during initial drafts of this chapter.

We would like to thank the Zang Lab for Computational Biology at the University of Virginia for their support.

References

Alexandrov, T., 2020. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. Annual Review of Biomedical Data Science 3. https://doi.org/10.1146/annurev-biodatasci-011420-031537.

Anscombe, F.J., 1973. Graphs in statistical analysis. American Statistician 27 (1), 17–21. https://doi.org/10.1080/00031305.1973.10478966.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Batarseh, F.A., Freeman, L., Huang, C.-H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 60. https://doi.org/10.1186/s40537-021-00445-7.

Belle, V., Papantonis, I., 2020. Principles and practice of explainable machine learning. arXiv:2009.11698 [cs,stat]. arXiv:2009.11698. Retrieved October 1, 2020, from http://arxiv.org/abs/2009.11698.

Belsley, D.A., 2004. Regression Diagnostics. Wiley Series in Probability and Statistics. Wiley, New York.

Bhattacharjee, R., Saini, L.M., 2015. Robust technique for the detection of Acute Lymphoblastic Leukemia. In: 2015 IEEE Power, Communication and Information Technology Conference (PCITC), pp. 657–662.

Bhattacharyya, G.K., Johnson, R.A., 1977. Statistical Concepts and Methods, 1st ed. Wiley. Retrieved July 20, 2018, from https://www.wiley.com/en-us/Statistical+Concepts+and+Methods-p-9780471072041.

Blackith, R.E., 1971. Multivariate Morphometrics. Academic Press, London.

Bolstad, W.M., 2012. Understanding Computational Bayesian Statistics, 1st ed. Wiley Series in Computational Statistics. Wiley, Hoboken.

Bravo González-Blas, C., Quan, X.-J., Duran-Romaña, R., Taskiran, I.I., Koldere, D., Davie, K., et al., 2020. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Molecular Systems Biology 16 (5), e9438. https://doi.org/10.15252/msb.20209438.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 2002. Manual on Setting up, Using, and Understanding Random Forests V3.1.

Caicedo, J.C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., et al., 2017. Data-analysis strategies for image-based cell profiling. Nature Methods 14 (9), 849–863. https://doi.org/10.1038/nmeth.4397.

Cleveland, W.S., 1993. Visualizing Data, 1st edition. Hobart Press, Murray Hill, N.J, Summit, N.J.

Costa, L.d.F., Roberto Marcond Cesar, J., Roberto Marcond Cesar, J., 2018. Shape Classification and Analysis: Theory and Practice, second edition.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., et al., 2011. Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. In: Special Issue on the 2010 Conference on Medical Image Computing and Computer-Assisted Intervention. Medical Image Analysis 15 (5), 729–737. https://doi.org/10.1016/j.media.2011.05.007.

Ding, Y., Yang, Y., Cui, Y., 2019. Deep learning for classifying of white blood cancer. In: Gupta, A., Gupta, R. (Eds.), ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging. In: Lecture Notes in Bioengineering, pp. 33–41.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, third edition. Wiley-Interscience, New York.

Duggal, R., Gupta, A., Gupta, R., Wadhwa, M., Ahuja, C., 2016. Overlapping cell nuclei segmentation in microscopic images using deep belief networks. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing. ICVGIP '16, pp. 1–8.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7 (1), 1–26. Retrieved July 2, 2018, from http://www.jstor.org/stable/2958830.

Euclid, 1728. Euclid's Elements. London.

Flusser, J., Suk, T., 1994. Affine moment invariants: a new tool for character recognition. Pattern Recognition Letters 15 (4), 433–436. https://doi.org/10.1016/0167-8655(94)90092-2.

Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36 (4), 193–202. https://doi.org/10.1007/BF00344251.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, 2nd ed. Chapman and Hall/CRC.

Gonzalez, R.C., Woods, R.E., Eddins, S.L., 2009. Digital Image Processing Using MATLAB, 2nd ed. by Rafael C. Gonzalez, 2nd edition. Gatesmark Publishing, S.I.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al., 2018. Recent advances in convolutional neural networks. Pattern Recognition 77, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013.

Gupta, R., Mallick, P., Duggal, R., Gupta, A., Sharma, O., 2017. Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple Myeloma. In: 16th International Myeloma Workshop New Delhi, India March 1-4, 2017. Clinical Lymphoma Myeloma and Leukemia 17 (1, Supplement), e99. https://doi.org/10.1016/j.clml.2017.03.178.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 26.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference 1988, pp. 23.1–23.6.

Hastie, T., Robert, T., Jerome, F., 2017. Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd (corrected 12th printing). Springer. Retrieved March 28, 2018, from https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21 (9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. John Wiley & Sons, Incorporated, New York, UNITED STATES. Retrieved April 7, 2018, from http://ebookcentral.proquest.com/lib/gmu/detail.action?docID=1138225.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24 (7), 498–520. https://doi.org/10.1037/h0070888.

Hu, M.-K., 1962. Visual pattern recognition by moment invariants. I.R.E. Transactions on Information Theory 8 (2), 179–187. https://doi.org/10.1109/TIT.1962.1057692.

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning. Springer Texts in Statistics. Springer New York, New York, NY.

James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics. Springer, New York. OCLC: ocn828488009.

Jiang, N., Burger, A., Crooks, A.T., Kennedy, W.G., 2020. Integrating social networks into large-scale urban simulations for disaster responses. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoSpatial Simulation, pp. 52–55.

Kaggle, 2017. Statoil/C-CORE Iceberg Classifier Challenge. Library Catalog: www.kaggle.com. Retrieved April 10, 2020, from https://kaggle.com/c/statoil-iceberg-classifier-challenge.

Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23 (3), 187–200. https://doi.org/10.1007/BF02289233.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement 20 (1), 141–151. https://doi.org/10.1177/001316446002000116.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., et al., 2018. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. Cell 174 (6), 1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039.

Kinser, J.M., 2018. Image Operators: Image Processing in Python, 1st ed. CRC Press, Boca Raton, FL.

Klinkenberg, B., 1994. A review of methods used to determine the fractal dimension of linear features. Mathematical Geology 26 (1), 23–46. https://doi.org/10.1007/BF02065874.

Kohavi, R., 1995. A study of cross validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, vol. 7. Retrieved May 14, 2018, from https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf.

Kulhalli, R., Savadikar, C., Garware, B., 2019. Toward automated classification of B-acute lymphoblastic leukemia. In: Gupta, A., Gupta, R. (Eds.), ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging. In: Lecture Notes in Bioengineering, pp. 63–72.

Labati, R.D., Piuri, V., Scotti, F., 2011. All-IDB: the acute lymphoblastic leukemia image database for image processing. In: 2011 18th IEEE International Conference on Image Processing, pp. 2045–2048.

Lamberti, W.F. [W.F.], 2020a. Classification of synthetic aperture radar images of icebergs and ships using random forests outperforms convolutional neural networks. In: 2020 IEEE Radar Conference (RadarConf20), pp. 1–6. ISSN: 2375-5318.

Lamberti, W.F. [William F.], Kinser, J.M., Kennedy, W.G., Eagle, M., Holmes, D.I., 2021. SVM-based models for pill shape classification. In: SDSS 2021. Retrieved from https://ww2.amstat.org/meetings/sdss/2021/onlineprogram/AbstractDetails.cfm?AbstractID=309635.

Lamberti, W.F. [William Franz], 2020b. Algorithms to Improve Analysis and Classification for Small Data. Ph.D. George Mason University, United States – Virginia. ISBN 9798557033350. Retrieved March 4, 2021, from http://search.proquest.com/docview/2476825035/abstract/90CC4207B46B4068PQ/1.

Lamberti, W.F. [William Franz], 2020c. Pill Shape Classification using Imbalanced Data with Human-Machine Hybrid Explainable Model. United States Patent Office: Application.

Lamberti, W.F. [William Franz], 2022. Classification of White Blood Cell Leukemia with Low Number of Interpretable and Explainable Features. arXiv:2201.11864 [cs, eess]. arXiv:2201.11864. Retrieved April 1, 2022, from http://arxiv.org/abs/2201.11864.

Laskey, K., Martignon, L., 2014. Comparing Fast and Frugal Trees and Bayesian Networks for Risk Assessment.

Lattin, J.M., Carroll, J.D., Green, P.E., 2003. Analyzing Multivariate Data. Google-Books-ID: VXxuQgAACAAJ. Thomson Brooks/Cole.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D., 1990. Handwritten digit recognition with a back-propagation network. In: Touretzky, D.S. (Ed.), Advances in Neural Information Processing Systems 2. Morgan-Kaufmann, pp. 396–404. Retrieved April 19, 2019, from http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf.

Li, X., Huang, W., Peters, D.K., Power, D., 2019. Assessment of synthetic aperture radar image preprocessing methods for iceberg and ship recognition with convolutional neural networks. In: 2019 IEEE Radar Conference (RadarConf), pp. 1–5. ISSN: 2375-5318.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., et al., 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society 389 (3), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x.

Lopes, R., Betrouni, N., 2009. Fractal and multifractal analysis: a review. Medical Image Analysis 13 (4), 634–649. https://doi.org/10.1016/j.media.2009.05.003.

Lundberg, E., Borner, G.H.H., 2019. Spatial proteomics: a powerful discovery tool for cell biology. Nature Reviews. Molecular Cell Biology 20 (5), 285–302. https://doi.org/10.1038/s41580-018-0094-y.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9 (86), 2579–2605. Retrieved May 19, 2021, from http://jmlr.org/papers/v9/vandermaaten08a.html.

Makhov, D., Samorodov, A., Slavnova, E., 2020. Different approaches for automatic nucleus image segmentation in fluorescent in situ hybridization (FISH) analysis for HER2 status assessment. In: 2020 26th Conference of Open Innovations Association (FRUCT), pp. 270–277. ISSN: 2305-7254.

Mancl, L.A., DeRouen, T.A., 2001. A covariance estimator for GEE with improved small-sample properties. Biometrics 57 (1), 126–134. https://doi.org/10.1111/j.0006-341X.2001.00126.x.

Manly, B.F.J., 1994. Multivariate Statistical Methods: A Primer, 2nd ed. Chapman and Hall, London.

Marzahl, C., Aubreville, M., Voigt, J., Maier, A., 2019. Classification of leukemic B-lymphoblast cells from blood smear microscopic images with an attention-based deep learning method and advanced augmentation techniques. In: Gupta, A., Gupta, R. (Eds.), ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging. In: Lecture Notes in Bioengineering, pp. 13–22.

McInnes, L., Healy, J., Melville, J., 2020. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [cs, stat]. arXiv:1802.03426. Retrieved April 29, 2021, from http://arxiv.org/abs/1802.03426.

Mendenhall, W., Sincich, T.T., 2011. A Second Course in Statistics: Regression Analysis, 7 edition. Pearson, Boston, MA.

Mikolajczyk, A., Grochowski, M., 2018. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp. 117–122.

Miller, E.G., Matsakis, N.E., Viola, P.A., 2000. Learning from one example through shared densities on transforms. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), Vol. 1, pp. 464–471.

Miller, R.G., 1974. The jackknife–a review. Biometrika 61 (1), 1–15. https://doi.org/10.2307/2334280.

Molnar, C., 2020. Interpretable Machine Learning. Google-Books-ID: jBm3DwAAQBAJ.Lulu.com.

Morency, C., Chapleau, R., 2003. Fractal geometry for the characterisation of urban-related states: greater Montreal case. Harmonic and Fractal Image Analysis, 30–34.

Murphy, J., Devereaux, A., Goodman, N.P., Koppl, R., 2021. Expert failure and pandemics: on adapting to life with pandemics. Cosmos + Taxis 9 (5+6). Retrieved June 7, 2021, from https://papers.ssrn.com/abstract=3773846.

Myers, R.H. (Ed.), 2010. Generalized Linear Models: With Applications in Engineering and the Sciences, 2nd ed. Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J. OCLC: ocn426796752.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015, pp. 41.1–41.12.

Pärnamaa, T., Parts, L., 2017. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. G3: Genes, Genomes, Genetics 7 (5), 1385–1392. https://doi.org/10.1534/g3.116.033654.

Peck, R., Olsen, C., Devore, J.L., 2008. Introduction to Statistics and Data Analysis, 3rd edition. Cengage Learning.

Plotze, R.d.O., Falvo, M., Pádua, J.G., Bernacci, L.C., et al., 2005. Leaf shape analysis using the multiscale Minkowski fractal dimension, a new morphometric method: a study with Passiflora (Passifloraceae). Canadian Journal of Botany; Ottawa 83 (3), 287–301. Retrieved July 24, 2020, from http://search.proquest.com/docview/218604942/abstract/F312523877A24D08PQ/1.

Preisser, J.S., Qaqish, B.F., 1996. Deletion diagnostics for generalised estimating equations. Biometrika 83 (3), 551–562. https://doi.org/10.1093/biomet/83.3.551.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 234–241.

Rosenfeld, A., 1974. Compact figures in digital pictures. IEEE Transactions on Systems, Man and Cybernetics SMC-4 (2), 221–223. https://doi.org/10.1109/TSMC.1974.5409121.

Russ, J.C., 1995. The Image Processing Handbook, 2nd ed. CRC Press, Boca Raton.

Sahlol, A.T., Abdeldaim, A.M., Hassanien, A.E., 2019. Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm. Soft Computing 23 (15), 6345–6360. https://doi.org/10.1007/s00500-018-3288-5.

Sahlol, A.T., Kollmannsberger, P., Ewees, A.A., 2020. Efficient classification of white blood cell leukemia with improved swarm optimization of deep features. Scientific Reports 10 (1), 2536. https://doi.org/10.1038/s41598-020-59215-9.

Samek, W., Wiegand, T., Müller, K.-R., 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. In: Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services. ITU Journal: ICT Discoveries 1. arXiv:1708.08296. Retrieved June 28, 2021, from http://arxiv.org/abs/1708.08296.

Shamir, L., 2011. Ganalyzer: a tool for automatic galaxy image analysis. The Astrophysical Journal 736 (2), 141. https://doi.org/10.1088/0004-637X/736/2/141.

Singhal, V., Singh, P., 2014. Local binary pattern for automatic detection of acute lymphoblastic leukemia. In: 2014 Twentieth National Conference on Communications (NCC), pp. 1–5.

Singhal, V., Singh, P., 2016. Texture Features for the Detection of Acute Lymphoblastic Leukemia, pp. 535–543.

Takei, Y., Yun, J., Zheng, S., Ollikainen, N., Pierson, N., White, J., et al., 2021. Integrated spatial genomics reveals global architecture of single nuclei. Nature 590 (7845), 344–350. https://doi.org/10.1038/s41586-020-03126-2.

Tibshirani, R., 1994. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Trzaskoma, P., Ruszczycki, B., Lee, B., Pels, K.K., Krawczyk, K., Bokota, G., et al., 2020. Ultrastructural visualization of 3D chromatin folding using volume electron microscopy and DNA in situ hybridization. Nature Communications 11 (1), 2120. https://doi.org/10.1038/s41467-020-15987-2.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science. Addison-Wesley PubCompany, Reading, Mass.

Union, E., 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Legislative Body: EP, CONSIL. Retrieved March 17, 2021, from http://data.europa.eu/eli/reg/2016/679/oj/eng.

Valen, D.A.V., Kudo, T., Lane, K.M., Macklin, D.N., Quach, N.T., DeFelice, M.M., et al., 2016. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. PLoS Computational Biology 12 (11), e1005177. https://doi.org/10.1371/journal.pcbi.1005177.

Wackerly, D., Mendenhall, W., Scheaffer, R.L., 2008. Mathematical Statistics with Applications, 7th edition. Thomson Brooks/Cole, Belmont, CA.

Wand, H., Lambert, S.A., Tamburro, C., Iacocca, M.A., O'Sullivan, J.W., Sillari, C., et al., 2021. Improving reporting standards for polygenic scores in risk prediction studies. Nature 591 (7849), 211–219. https://doi.org/10.1038/s41586-021-03243-6.

Wang, J., Mall, S., Perez, L., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621. 8.

Way, G.P., Greene, C.S., 2018. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pacific Symposium on Biocomputing 23, 80–91.

Wickham, H., 2016. Ggplot2: Elegant Graphics for Data Analysis, 2nd ed. 2016 edition. Springer, New York, NY.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann.

Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., et al., 2019. Genetic analyses of diverse populations improves discovery for complex traits. Nature 570 (7762), 514–518. https://doi.org/10.1038/s41586-019-1310-4.

Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T., 2019. Classification-Reconstruction Learning for Open-Set Recognition, p. 10.

Zakrzewski, F., de Back, W., Weigert, M., Wenke, T., Zeugner, S., Mantey, R., et al., 2019. Automated detection of the HER2 gene amplification status in Fluorescence in situ hybridization images for the diagnostics of cancer tissues. Scientific Reports 9 (1), 8231. https://doi.org/10.1038/s41598-019-44643-z.

Zeiler, M.D., Fergus, R., 2013. Visualizing and understanding convolutional networks. arXiv:1311.2901. Retrieved April 15, 2019, from http://arxiv.org/abs/1311.2901.

Zeng, X., Cao, K., Zhang, M., 2017. MobileDeepPill: a small-footprint mobile deep learning system for recognizing unconstrained pill images. In: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. MobiSys '17, pp. 56–67.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: a 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009.

Zode, J.J., Choudhari, P.C., Uparkar, M., 2017. Comparative study of methods to determine fractal dimension. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pp. 441–446.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, Statistical Methodology 67 (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

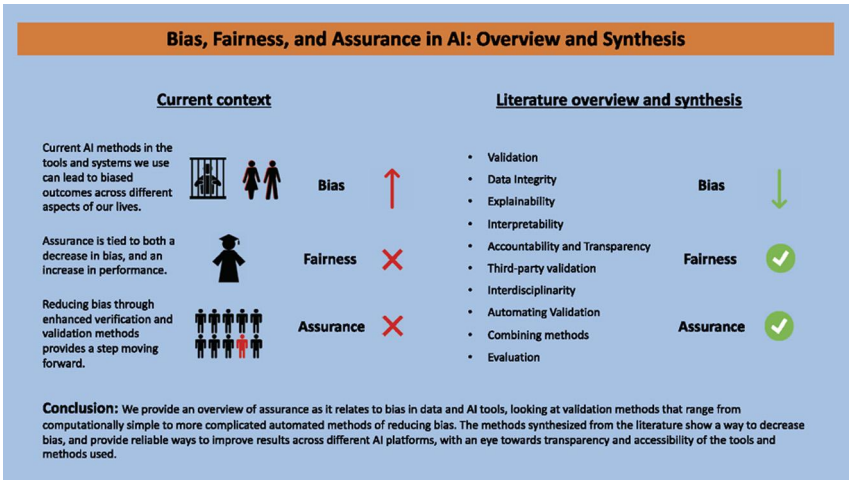
This page intentionally left blank

4

Bias, fairness, and assurance in AI: overview and synthesis

Amira Al-Khulaidy Stine and Hamdi Kavak
George Mason University, Computational and Data Sciences Department, Fairfax, VA, United States

Graphical abstract



Abstract

Artificial Intelligence (AI) has provided one of the most significant breakthroughs in computing in the past several decades. From the automated landing of space rovers on the martian surface and self-driving delivery trucks to making swift trading decisions, AI has significantly changed how we perceive intelligent computers or robots. However, all these advancements in AI and their real-world applications come with an increased cost of validation. Validation in AI is an extensive topic dealing with challenges related to assessing safety and thorough usability. The advances and widespread use of AI come with challenges in validating AI, specifically for bias, fairness, and assurance. On the positive side,

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00016-0

many studies in the past decade have focused on such challenges concerning technical, legal, ethical, and philosophical aspects. Now, it is time to look at these AI challenges with more holistic eyes.
In this chapter, we first provide a comprehensive introduction to the bias, fairness, and assurance concepts in AI, starting with their definitions. Then, we discuss how AI has had unintended effects on society and present some concrete examples where this occurs. We then provide an overview of the literature that covers validation methods for AI from different aspects. Improved AI assurance and understanding of how bias occurs is important, as well as tackling the issue of bias in model validation. By synthesizing the literature, this chapter provides an overview of ways to formalize the process of bias reduction and assurance in AI. This comprehensive synthesis provides AI researchers, data scientists, policymakers, and practitioners a way to assess how a particular AI model can be evaluated against bias, fairness, and assurance for a particular goal.

Keywords

AI assurance, bias in AI, AI validation, AI fairness

Highlights

- AI systems can support better decision-making
- Current AI platforms can lead to biased outcomes
- AI assurance is tied to both a decrease in bias and an increase in performance
- Enhanced validation methods constitute one of the means of facilitating AI assurance
- A synthesized set of methods provides a reliable way to improve results across different AI platforms

4.1 Introduction

Artificial Intelligence (AI) is no longer a futuristic notion but a feature of everyday life. Technological advances have allowed AI to be in many forms, from algorithms, natural language processing (NLP) to robots is present in fields from education and medicine to social media and shopping (Blodgett et al., 2020; Hagras, 2018; Osoba and Welser, 2017). The abundance of algorithms that exist to take in large amounts of data to facilitate processes

and systems has become the norm for how governments and businesses operate (Munoko et al., 2020).

Our reliance on AI systems in almost every aspect of our daily lives has made it more apparent how much bias has been translated into the AI systems that we depend on. The day-to-day need for transparency in AI systems and innovative techniques for how we can validate our systems for fairness and assurance is crucial to the performance of our systems and the reduction of injustices that arise from biased tools. This chapter is meant to provide a synthesis and an overview of validation methods that can be useful for AI researchers, policymakers, and practitioners from all fields, so that these methods can be incorporated into the systems we create and use. Perhaps, what we can do is not exacerbate the problem, but instead ensure that we can maximize the objectivity of our AI tools to perform with less bias and higher efficiency than humans do at these tasks for which the AI has been built.

Mathematical models have become the guidepost for decision-making and "micromanaging the economy, from advertising to prisons" (O'Neil, 2016). Assurance in AI is a growing field, meant to tackle the issue of bias and errors with machine learning methods and AI, as well as ensuring that accuracy is attained with the same or greater level of expertise that a human would have in that role (Fujii et al., 2020; Dwork and Ilvento, 2018), "including outcomes that are valid, trustworthy, and ethical, unbiased in its learning and fair to its users" (Batarseh et al., 2021).

By building AI systems and models that incorporate these elements into the outcome, the task that belonged to an error-prone human can be optimized for better performance and reduced bias. Many other complications arise from organizations and institutions relying on AI outcomes, where automated systems make recommendations that are either ignored due to "algorithm aversion" and a lack of trust in the outcomes provided by the AI system, or at the other extreme, blindly employed, despite information showing that the AI output is incorrect (De-Arteaga et al., 2020).

As more data becomes available for training AI systems in different fields, ethical and legal issues regarding human data, privacy, and equity come to the forefront. Achieving a balance between the uses of AI and the protection

of individual privacy while remaining within the realm of legality and ethics is a rising challenge (Rodrigues, 2020). Though testing of AI has shown advances, particularly with the availability of data (Byun and Rayadurgam, 2020; Gore et al., 2017), the issue of bias within data and within our social constructs and their reproduction in AI is still an ongoing issue, particularly when relying on statistical output without an understanding of the underlying mechanisms that produce these statistics (Lum and Isaac, 2016). To this point, there is a difference between "statistical fairness, and individual fairness" (Chouldechova and Roth, 2018), where minimizing the average error to fit the populations has extreme results on the fairness at the level of the individual. This is an example of bias in AI, where the biased data and the need to minimize error for a better fit have ethical consequences.

How do we ensure that our AI systems are tested to maximize performance while also ensuring results that comply with ethical frameworks? Can we ensure that in our validation efforts in AI, we can comply with certain parameters and agreed-upon standards for minimizing bias in AI? Assurance has traditionally been understood within the context of something performing well in both reliability and fairness (Batarseh et al., 2021; Ören, 1987), but is it also being accountable? This can be the case in fields where decision-making supported by AI can lead to human harm, such as in the medical field, or where finances and audits, or prison sentences are concerned (Dressel and Farid, 2018; Munoko et al., 2020; Habli et al., 2020).

Beyond some of the theoretical and philosophical underpinnings of these questions, this chapter seeks to provide an overview and compilation of the current methods that allow for processes to be examined and tested to ensure that the algorithms we use perform within an ethical framework that can add to existing methods of performing assurance (Balci, 1997).

4.2 Assurance and ethical AI

To understand how AI can be ethical, we could turn to a more philosophical notion of ethics, but discussions over the ontology of ethics in AI would take away from the task since what is needed is a concrete and transparent way of understanding ethical AI. Particularly given the practical applications of AI and the need for a common understanding of how to measure

AI assurance, as well as for purposes of testing, a more pragmatic definition is needed to ensure that the goal of fairness and decreased bias is attained. For this, we will reduce the focus on the philosophical discussion of ethics in AI to provide a more concrete definition for how we can interpret what is ethical and what is not. At the basis for our understanding of ethical AI, we rely on the existing legal protections that are formed by social contract and the law (McNamara et al., 2017).

These legal protections cover discrimination against race, gender, ability, and many other attributes, and we assume those to be the pillars on which the ethical performance of AI is measured against. AI bias can be seen when two individuals with the same attributes, with differing protected attributes, receive different decisions from an AI program. An example of this would be two individuals having identical attributes, with the exception that one individual is male and the other female, or where two individuals have the exact attributes with only "race" being the differing attribute (Romanov et al., 2019; Agarwal et al., 2018).

4.2.1 Overview of bias and lack of assurance in AI

The reliance of AI on a single attribute or a set of attributes that do not have any relevance to the output is a feature of a system that relies on available data, but cannot discern, qualitatively, which attributes provide relevant outputs. This reliance on data without the next step of having knowledge of what is predicted is central to the question of bias. One such example where this is clearly demonstrated is with the issue of recidivism and prison sentencing. Models that predict recidivism have been shown to favor individuals who fall within certain races, and in Section 4.2.2, we give a more detailed example of how this occurs. By providing data of biased human decisions from previous cases to the AI tool, the data is then used to maintain a loop, where individuals are not able to receive fair sentencing due to prevalent discrimination in the legal systems that support sentencing decisions (Nabi and Shpitser, 2018). Other examples show that college acceptances and placements favor individuals with certain profiles and greatly disfavor individuals who do not meet a certain criterion that has little to do with performance and much to do with race, gender, and socioeconomic

status (Osoba and Welser, 2017). The challenges have traditionally been consigned to policy issues, but with AI replacing many human-led systems, where individuals were the decision-makers, the need for greater AI assurance to optimize these decisions becomes crucial.

Many examples of bias in AI come from discrimination based on race, gender, socioeconomic status, and educational attainment, because surveys usually contain fields for these types of attributes, and the prevalence of bias from these attributes occurs so widely and across so many problem spaces, that they are easier to spot. Some other ways where bias occurs and has gained significant attention comes from using names in job applicant selection (Carlsson and Rooth, 2008), social media user detection (Mislove et al., 2011), as well as determining disparities based on race and ethnicity (Elliott et al., 2009). Names, which can, in turn, be associated with gender, ethnicity, and race, lead to a new way of using data to maintain bias through an attribute that identifies a person, but does not provide any true context beyond being a vehicle for biased decision-making. This is another example of how human biases are introduced into the AI systems used that end up creating an unreliable biased tool that may be faster than a human, but is just as unreliable.

Data bias can be due to poor sampling methods, historical biases based on socioeconomic factors, or decisions on categorizing and defining data. Mehrabi et al. (2019) provide a comprehensive list of different types of AI bias and examples of how they occur. Data bias can also be due to representation bias, measurement bias, and the choices of how to measure an attribute. Aggregation and population bias have to do with assumptions resulting from observations of a certain population. Simpson's paradox is a good example of aggregation bias, where some conclusions about subsets of the data which are meaningful and informative become erased, reversed, or changed when the subgroup is considered within the context of the entire sample (Malinas and The Hegeler Institute, 2001). Due to the aggregation of the data, the data provides conclusions about the population as a whole, and any other conclusions about subsets of the data become absorbed and disappear into the larger numbers of the entire population.

Behavioral bias is another type of bias that relates to how users behave in different contexts and platforms that come along with social media use. Some ways behavioral bias can be observed in social media use can be broken down into popularity bias, ranking bias, and other human-related behaviors that are taken to be representative, despite the data being highly contextual to certain situations and platforms (Olteanu et al., 2019). These are some of the ways in which bias occurs and how pervasive bias can be, highlighting how difficult it is to detect bias in many situations, including our use of these tools in our daily lives.

Thus far, we have maintained a fundamental assumption in this chapter that AI should be less biased and more objective than a human in a similar role. We maintain this because technology has traditionally been implemented to improve upon human systems that have been inefficient and perhaps ineffective (Varshney, 2019). With AI playing a greater role in all processes, there is an underlying understanding that the efficiency and effectiveness of AI should outperform that of a human or allow for processes to become easier for humans to follow (Madras et al., 2018). This can be shown by improvements in medicine, such as automating parts of surgery. However, one issue that continues to be present is that AI systems are heavily reliant on data (Hagras, 2018), and while we have a lot of data available, the integrity of both the sampling methods, the cleanliness of the data, as well as issues of privacy and ethical uses of the data become a factor.

Data, due to collection methods, interpretations, categorizations, and other decisions that need to be made for the data to be useful, comes with some built-in bias (Pedreshi et al., 2008). Many times the reading in of data is automated, and the next system follows along with the next steps with less emphasis on verifying and validating the data or other aspects of the AI system. In Section 4.3, we discuss how this process of validating our data is also an important step in AI assurance, and we provide some suggestions for how this can be done after we share some examples and limitations of how this bias manifests in different situations.

4.2.2 Current assurance methods for bias reduction

Goodhart's law states that "when a measure becomes a target, it ceases to be a good measure" (Strathern, 1997). This has proven true in standardized testing, link recommendation algorithms, and medical software (Shore and Wright, 2015). In the example of incarceration and recidivism, the attempt to predict which individuals would commit crimes and return to jail proved to be a self-fulfilling prophecy. The measures were over-emphasized at the expense of the true target, leading to a perpetuation of systemic racism and bias (Chouldechova, 2017; Dressel and Farid, 2018; Mahadevaiah et al., 2020). As the AI trains on a biased sub-set of data, which already includes biases, a cyclical process develops, where meeting measurements provided by the data is the goal. In social media sites, this can be observed through echo chambers, where there is very little cross-over in information sharing and a large presence of polarization (Papakyriakopoulos et al., 2020). This occurs because AI is rewarded with validation against biased data, and it does not go beyond that.

An example in the field of standardized testing is to favor students who meet scores on specific exams (Ntoutsi et al., 2020). The end goal is to demonstrate knowledge of the subject matter and translate learning skills from one educational setting to another. However, what occurs instead is that receiving a high score on the test becomes the goal, regardless of whether there is sufficient comprehension of what is being tested. With the process of automated scoring for these exams, as well as systems that screen applicants to educational institutions, the human element, which may have had more context for interpreting the applicant holistically, is removed, and instead what is left are students who match one profile, decreasing diversity and representation, but increasing bias in a domino-like effect, which is then perpetuated in the job market, and other social spheres (Zemel et al., 2013).

This is also the case in recruiting for large companies, where employees may never be vetted due to the bias in the AI systems used for screening applications (Abdollahi and Nasraoui, 2018). In some ways, the optimization of the work that was performed by a human fails to meet assurance stan-

dards, because the AI system simply becomes a faster stand-in for the same flawed process it was designed to replace.

Chen et al. (2018) describe fairness as something which can be evaluated in the context of protected groups. A marketing model that includes bias and ends up targeting based on certain characteristics may be less harmful than a model that is meant to save lives in the health field (Habli et al., 2020) or provides decision-making assistance to lawmakers. Determining what is fair is a philosophical problem, but some definition is important to practically tackle this issue. There is also the greater issue of a trade-off between fairness and accuracy, particularly in reinforcement learning, but this trade-off can be reduced. There are simple but effective solutions for doing this, which entail increasing the training set size, and measuring additional variables that may result in less issues with bias (Chen et al., 2018). Chen et al. (2018) also illustrate this with some examples on income prediction, and mortality predictions from clinical notes, and find that bigger training sets, and looking at other variables, help to greatly reduce the issue with fairness, while still keeping accuracy.

Across the AI community, this is also a question that comes up as an ongoing balance and trade-off between exploration and experimentation of machine learning models. Exploration is the process of training the model with data (which generally contains bias) and experimentation, where these models are then tested, usually with users participating in this process by organically using tools, such as social media applications (Bird et al., 2016). Bird et al. (2016) state that both of these methods present ethical and social issues, including privacy concerns and how to navigate a field and methodology that moves so quickly in innovation that it is difficult for policies and laws to keep up. This is an ongoing issue for the population as a whole, with different complexities that are difficult to capture and tackle fully.

To further expand on a specific example we have mentioned, Chouldechova (2017) looks at recidivism prediction instruments (RPI's), which provide decision-makers with an assessment of whether a criminal defendant is more or less likely to commit a crime in the future. Chouldechova (2017) illustrates that the issue with some of these instruments is that they make

into a statistical problem, a problem that is essentially ethical. Chouldechova (2017) uses the example of Boward county data and a tool called the correctional offender management profiling for alternative sanctions (COMPAS) RPI, which relies on a psychometric test to look at two races, Black and Caucasian, and predicts on a few attributes, such as previous crimes and severity of crimes the "probability of recidivism." The instrument uses a classifier based on a threshold, as well as the recidivism prevalence between groups. The outcome of the RPI shows that recidivism of Black defendants was 51 percent versus 39 percent for White defendants, meaning that the tool determined, based on race alone and all other attributes being equal, that 51 percent of people who had their race listed as Black were more likely to commit crimes after being released. The results for every range of prior offenses (regardless of whether there were 0 or more than 10) show that Black defendants are categorized as being more likely to commit future crimes based on the RPI. How is this possible?

One explanation is that the distribution used for the underlying RPI instruments, as well as the use of binary parameters, show that the output is heavily biased towards White defendants being granted bail more often. One of the reasons that White defendants are being granted bail more often is exactly because the data going into these decision-making tools is biased towards sentences that grant bail more often to White defendants. The research focused on race and crime show that there is great racial disparity in the U.S. criminal justice system, especially when looking at the percentages of a minority population in society compared to the percentages of minority populations that are represented in the criminal justice system (The Sentencing Project, 2018; Taxman and Byrne, 2005). These disparities regarding race and ethnicity are present in the data, which leads to systems and tools using the same biased "decision-making" that prove to be just as biased and unreliable as human decision-makers.

Despite the possible unreliability of these tools, such as in the example of the COMPAS RPI, these tools can still be more useful and error-free than human judgment, particularly when implemented with an understanding of assurance in mind. However, it is important to illustrate how a model can be rendered unusable due to the self-enforcing loops of biased sentences.

These sentences form the basis of the data that is meant to train the system to determine whether defendants are more likely to commit crimes, but due to being biased data, the AI system performs poorly in terms of fairness and assurance. These models are meant to be set up to contain the correct attributes, but how we interpret the results concerning the error scores, confidence intervals, and the constraints on the false positive and false negative rates depends on the user. However, if bias is present in the data, the model works against the intended use, and the measures it provides could end up doing more harm than good (Dressel and Farid, 2018).

Malik (2020) provides another definition of limitations of machine learning by creating a hierarchy of limitations and focusing on four key aspects. Beyond the general limitations of models, summarized by George E Box as "all models are wrong, but some are useful" (Box, 1976), there are other limitations beyond usefulness and the lack of complete representation that we should address when modeling. The four key aspects are:

1. A reliance on using only quantitative analysis over qualitative analysis, when a combination may be more thorough;
2. Using probabilistic modeling instead of mathematical models or simulations;
3. Focusing on predictive modeling over exploratory modeling;
4. Reliance on cross-validation, in lieu of evaluating model performance.

These issues all contain limitations that leave us with the following problems: the first issue is leaving the qualitative parts of the problem space aside to the detriment of our solution-space. By doing this, we are left with a loss in significance, context, and meaning. This point asks us to focus on the question at the very heart of modeling, which is what is being asked of the model and our problem space. The measurements and responses are context-dependent, so a qualitative explanation with descriptions of the problem space needs to be included in both the formulation of the model as well as the validation of the model (Malik, 2020; Brennen, 2020).

The next issue is probability; by using probability, we force outliers and entities to fit a structure that produces a minimization in error. Many times this minimization contributes to the bias and loss of fairness by forcing a central tendency, where one may not naturally exist. A focus on predictive

modeling takes correlations and attempts to make larger pronouncements about them, which may result in many issues and misinformation. Alternatively, by focusing on the exploratory analysis, we can understand some of the causality, which can be more helpful, and "change things [for better] in the world by looking at the data and results more holistically" (Malik, 2020).

The over-reliance on cross-validation, instead of using model performance, tends to be more forgiving of models that over-fit. A model that is based on performance may undergo more scrutiny and may take longer to become validated. Malik (2020) uses the medical field as an example, where a model is useful for how it does on actual tasks where accuracy is crucial. This leads to the greater question of how we can validate these models to represent the real world accurately and provide the least biased representation of what is being modeled.

Jacobs and Wallach (2019) show that the bias lies in the translation of what goes into the first stage of model building, under "construct" that is initially masked, and then is modified in various steps of the "operationalization" stage. Several steps along the way, at each stage of the modeling process, there can be errors resulting from biases that were abstracted from the initial biases that went into the model. The authors show some examples of this using the simple example of "height." The construct of height is seemingly simple and intuitive; however, when operationalized, several questions come up, such as position, "slouching," whether the person is in a wheelchair and other instruments that aid in measuring height, which may introduce a set of questions and issues that need to be looked at.

As with any representation, the authors emphasize that modeling these terms presents sub-questions that need to be decoded. Another example is socioeconomic status. "Income" can seem like the correct way to operationalization the construct of socioeconomic status, but it turns out that it is more complicated and requires taking into account many other factors as attributes that may not initially seem intuitive, including geographic area, the compared income of other people in that area, social standing, and so on. Thinking about the problem-space and social constructs and how we operationalize them becomes a key factor in ensuring that our AI is

validated throughout each process to reduce and avoid the introduction of bias.

We have previously discussed how validation methods are an important part in ensuring that our systems and models function appropriately for both the task they were built and outputting information that is both useful and fair. Validation is tied with assurance, in the sense that both refer to the output of the AI system and provide an answer to the question of whether the system or model "is the right tool, doing the right job" (Balci, 1997). Validation can be an extensive process, and there are several ways in which it is accomplished, as we will see in the next section.

4.3 Validation methods

The resulting question from this is how then we can measure bias, given that it presents in many different ways, and may not be something we may thoroughly know to consider. What are some ways we can work to make these models more "fair"? Bird et al. (2016), for example, present a solution by modifying a naïve Bayes classifier to be "discrimination-free." Other suggestions include modifying the probability distribution of the sensitive attribute to artificially remove the positive values from one class by adding more probability values to the discriminated sensitive values, which improves the "discrimination" value, but leaves all other attributes untouched (Romanov et al., 2019).

Another approach is to use the latent variable model, which forces the model to focus on only the attribute deemed to be the cause of the discrimination to optimize the parameters. These approaches presume prior knowledge of the bias, which in the case of gender-pay inequality may be simple (Calders and Verwer, 2010), but can be a problem with more attributes and how they relate.

In their example of income by gender, Calders and Verwer (2010) obtain a probability of male versus female earning more, and the results show there is income disparity based on gender. However, though this may be the result of their naive Bayes model based on gender, the authors bring up a counterpoint of how do we know if this is discrimination by the model or a reflection of the redlining effect, which is present in the world, where

pay discrimination based on gender truly does exist. Though they state that their approach does help solve the issue for models that are fairly simple and look at one attribute, it can become more complicated with several interrelated attributes and may result in other effects, which are unknown.

Measuring discrimination and fairness is challenging. Jacobs and Wallach (2019), provide another approach that focuses on the validation of the model throughout the modeling process. They propose using an array of different types of validation techniques to test our models. The authors use interdisciplinary methods of cross-validating a range of models to produce a few different ways to approach the issue of validating models for bias.

Though the focus has been largely on validating models to decrease bias, verification is also an important factor that allows for better decisions when partitioning data for training and testing, as well as all the effort that goes into the AI system before the user sees any results. We have mentioned some verification efforts, particularly when looking at ways to reduce bias in the input data. Verification is a crucial first step in reducing some of the bias by ensuring that the system is functioning correctly, with the correct inputs needed, and some new ways of automating some of the verification processes can prove to help increase assurance. Automating some of the verification and validation processes can help bolster efforts in AI assurance. However, regardless of the techniques used, the crucial part of AI assurance is to examine and calibrate our AI tools to reduce constantly, and if possible, eliminate bias (Balci, 1997; Wing, 2020; Lynch et al., 2020).

Optimizing the systems that do this demands time and effort and context knowledge, which we understand as the role of understanding the nuances of a problem that comes from understanding the context of the problem-space. This term is related to domain expert or third-party stakeholder (Srivastava and Rossi, 2019), which emphasizes that experience in understanding the problem area as a whole can allow for new insights, which may not be as easily gleaned from our AI systems (Dobson, 2015).

In studies where stakeholders were asked about what AI assurance meant to them, similar themes included a lack of transparency, primarily when used to understand the black-box decisions of many of the AI systems (Brennen, 2020). Difficulties in identifying bias and in utilizing the

decisions for analysis due to output that is not user-friendly. Working in interdisciplinary teams to create checklists and standards for ensuring a common understanding can result in a reduction in bias throughout the process (Madaio et al., 2020). Third-party validation of decisions and bias recognition, particularly in languages and language translation software (Srivastava and Rossi, 2019), can also support AI assurance and increase accountability.

Accountability and transparency come from reducing the opacity of AI systems (de Laat, 2018). Transparency could increase with a push for presenting AI decisions and outputs as ranked based on performance, as well as including a way to compare different models and output. Along with transparency comes the need to protect individual privacy. It is a difficult balance; however, starting with multi-model systems and comparing predictions, privacy can be maintained while still being judicious about using AI for decision-making.

Additionally, the high test accuracy can be seen as a favorable classifier outcome, but it can allow for hidden bias. With the introduction of new data and more scrutiny for how the high accuracy is achieved, and a better understanding of how to interpret and deal with noisy data, more trustworthy AI can be developed (Zhang et al., 2018; Fogliato et al., 2020; Go and Lee, 2018). An expanded version of utilizing data to bolster the AI system is to "harness adversarial examples" instead of consistently using and relying on representative data. Another approach relies on using "corrupted or inconsistent training data" to build a more robust sample and help make models less biased (Goodfellow et al., 2015; Kaul, 2018; Kulkarni et al., 2020).

Automated ways of validating AI systems and models can also help with speed while maintaining accuracy. Due to the extensive efforts required to validate AI systems, some authors propose automated methods that bolster validation by constructing accurate and effective unit tests that allow the system to be tested and improved in each step of the process (Breck et al., 2019). These proposed techniques are generally accompanied by decision output that is user-readable and explainable. This process allows for automation by creating better tests and allowing the user to have input on the decisions. Some examples proposed by automating the explainability

of predictions include using restricted Boltzmann machines (RBM), with a restriction that includes the formation of paired nodes from each group so that nodes become "visible" as they connect to another single-node, instead of falling within the "hidden" group, where some of the learning becomes obfuscated to the user, which has been the case with many neural networks (Abdollahi and Nasraoui, 2018).

Similarly, individual discrimination can be reduced by combining symbolic execution and local expandability and interpretability so that effective tests are generated, with the predictions as user-friendly output, such as decision trees or linear models. Other examples utilize a process for reducing data errors that "adversely affect the quality of the generated model," but provide a more nuanced solution space. This is accomplished by including a data analyzer, a data validator, and a model unit tester to perform checks along the way (Breck et al., 2019).

4.4 Synthesis of the literature

Many of the approaches are best used in conjunction with others, and the more thoroughly validated each step in the AI system is, the greater the chances of the final system providing reliable and fair outcomes (Yang and Stoyanovich, 2017). Jacobs and Wallach (2019) provide a breadth of definitions and examples of how to validate AI along each part of the process of utilizing AI. Due to the thoroughness of their work, we have combined their validation methods along with some of the other proposed bias-reducing methods into the following categories. Table 4.1 is a compiled list of the different types of validation methods, and additional examples we have summarized from different ways of reducing bias in AI and increasing assurance. The table provides validation techniques and ways to frame the problem space with an eye for bias. We provide a comprehensive list that seeks to capture ways we can implement these techniques in our current AI systems, particularly within the greater framework of verifying and validating the models we use.

Validation: Validation is an important step in bias-reduction. Jacobs and Wallach (2019) provide several categories and descriptions of how this can

be accomplished. Construct validity, which interrogates the quality of the measurement used; face validity, which asks if measurements produced look correct; content validity, which checks to see if the model captures everything it needs to; convergent and discriminant validity, which checks for models aligning with other models; predictive validity and hypothesis validity, which relates to the correct properties and theoretical constructs of the model; and consequential validity, which relates to the consequences of the outcomes. Each sub-category listed is meant to interrogate our tools to ensure a minimization of bias (De-Arteaga et al., 2020; Jacobs and Wallach, 2019; Mehrabi et al., 2019).

Data integrity: This category covers better partitioning of data for training and testing, as well as checking, quantitatively and qualitatively, that our system has the correct inputs. Data integrity also refers to the quality of the chosen measurements, as well as reducing data errors that "adversely effect the quality of the generated model." Some innovative ways of doing this with data include "harness[ing] adversarial examples" in our data, instead of consistently using and relying on representative data. This step also encourages questioning high test accuracy to ensure thoroughness of data being used (Jacobs and Wallach, 2019; Goodfellow et al., 2015; Balci, 1997; Ören, 1987).

Explainability: This category relates to the clarity of the model and outputs, as well as how the tool can be explained. Consulting with domain experts and third-party stakeholders can help us check if the model is usable and if anything is operating with assumptions that may be incorrect. Using multi-model systems instead of one system helps, as it offers the opportunity for several systems to produce results that can be compared to each other. Similarly, predictions from these tools can be ranked and explained for a more thorough understanding of the tool (De-Arteaga et al., 2020; Jacobs and Wallach, 2019).

Interpretability: A similar concept to explainability, but it differs in who the audience is. Interpretability has to do with the "black box" operations that occur, as well as the mechanics of the systems. Explaining the outputs and predictions is less important for this step than interpreting what is occurring with our AI tools and reducing the opacity of our AI systems. This step

helps reveal any assumptions that may be embedded and helps bring the inner workings of our tools to light (Jacobs and Wallach, 2019; Dwork and Ilvento, 2018).

Accountability and transparency: Accountability and transparency is a critical step in bias reduction. It provides the basis for questioning assumptions, as well as making the AI tools more usable and accessible to a wider public. This helps reduce "black box" decisions, but also forces accountability for the outcomes and predictions of the systems. It allows for users to question the system openly, as well as the ability to replicate the systems to test for biases (Batarseh et al., 2021; Jacobs and Wallach, 2019; Abdollahi and Nasraoui, 2018).

Third-party validation: Optimizing the systems with guidance from domain experts or third-party stakeholders who can provide more informed feedback. Reliability of using these tools in the real world with both users and testers, as well as subject-matter experts, policymakers, academics, and others who can agree on the systems and the implementation of the tools (Jacobs and Wallach, 2019; Srivastava and Rossi, 2019; Batarseh and Gonzalez, 2018; Chouldechova and Roth, 2018).

Interdisciplinarity: This encourages a wide range of fields to converge and allows for different fields to contribute to different parts of the system. This allows for a holistic and thorough approach in exploring a complete version of the system being created. AI tools are built on assumptions and ideas about how they should operate and their possible uses. By engaging experts from various domains, the tools can be more thoroughly, completely, and thoughtfully built (Madaio et al., 2020; Jacobs and Wallach, 2019; Batarseh and Gonzalez, 2018; Chouldechova and Roth, 2018).

Automating validation: There is consensus that validation is challenging work, and ensuring constant validation, verification, and calibration of our AI tools is difficult. Providing innovative and reliable ways of automating some of these steps ensures that validation is a priority during the entire process (Jacobs and Wallach, 2019; Breck et al., 2019; Abdollahi and Nasraoui, 2018; de Laat, 2018).

Combining methods: This category contains all the other categories and their processes. It is a thorough way of checking to make sure that more than one method of validation is included in our AI assurance process. It also encourages interrogating "corrupted or inconsistent training data" to build a more robust sample, instead of removing any noisy data points. Combining methods indicate that bias-reduction techniques are best used in conjunction with one another (Kulkarni et al., 2020; Breck et al., 2019; Goodfellow et al., 2015; Kaul, 2018).

Evaluation: Consistently testing and checking our systems for bias and being thoughtful about the effects of our tools. This relies on constructing accurate and effective unit tests that allow the system to be tested and improved in each step of the process. The evaluation of our tools allows users to have input on the decisions, as well as provide feedback to the tools through reinforcement learning. This step allows for users to actively engage with and interrogate the biases of the tools being used (Brennen, 2020; Jacobs and Wallach, 2019; Zhang et al., 2018; Agarwal et al., 2018; Messick, 1998).

We have also provided some suggested approaches, questions, and examples that we hope capture these different methods of reducing bias in AI. Table 4.1 is a compilation of the categories and is meant as a starting point for the expansion of additional methods of bias reduction. In this section, we would like to highlight that there is not a "one-size-fits-all" method for bias reduction. The methods used will depend highly on context and the problem space, but we would like to suggest that a thorough approach may be one where there is a combination of the different methods we explore.

4.5 Conclusion

In this chapter, we have provided a general introduction and overview of AI assurance. We have also explored some definitions of AI fairness, bias, and assurance, with some examples of how this occurs in different fields. We have also shared how AI assurance is currently being tackled in different domains and show some concrete ways of implementing different types of validation methods that lead to AI assurance.

Table 4.1 Approaches for bias reduction in AI.

Categories of bias reduction	Questions to ask and examples	Key references
Validation	- Have we interrogated our process throughout each step? - Do the measurements capture relevant facets? - Are the measurements valid at a glance?	Blodgett et al. (2020) De-Arteaga et al. (2020) Jacobs and Wallach (2019) Mehrabi et al. (2019)
Data integrity	- Do our measurements match other accepted measurements of this problem space? - Performing through cross-validation and other validation techniques.	Balci (1997) Jacobs and Wallach (2019) Ören (1987)
Explainability	- Is our AI explainable? - Do stakeholders and users have the ability to use the outputs for real-world problems? - Are we increasing accessibility and clarifying "black box" processes?	De-Arteaga et al. (2020) Jacobs and Wallach (2019)
Interpretability	- "When a measure becomes a target, it ceases to be a good measure." Goodhart's Law. Is this happening? - Are we using the data to gain insight or reaffirm biases? - Are we relying on the tool for answers, or is the AI a support to decision-making?	Dwork and Ilvento (2018) Jacobs and Wallach (2019) Strathern (1997)
Accountability and transparency	- What are the societal impacts of the AI tool? - What are some possible outcomes of relying on the outputs of our AI? - Are we able to explain the outcomes and interrogate possible biases?	Abdollahi and Nasraoui (2018) Batarseh et al. (2021) Jacobs and Wallach (2019)
Third-party validation	- Do we have a good theoretical understanding of the problem space? - Are we relying on users and subject-matter experts for thoroughness? - Ex: Using language experts for natural language processing and translation software	Batarseh et al. (2021) Chouldechova (2017) Jacobs and Wallach (2019) Srivastava and Rossi (2019)
Interdisciplinarity	- Are concepts represented accurately by all or many of their components? - Ex: Using language experts for natural language processing and translation software	Batarseh et al. (2021) Chouldechova (2017) Jacobs and Wallach (2019) Madaio et al. (2020)
Automating validation	- Is our AI producing reliable results consistently? - Does automatically updating the AI with new datasets produce reliable results? - Ex: Restricted Boltzmann machines (RBM) making each step visible	Abdollahi and Nasraoui (2018) Breck et al. (2019) de Laat (2018) Jacobs and Wallach (2019)

continued on next page

Table 4.1 (continued)

Categories of bias reduction	Questions to ask and examples	Key references
Combining methods	- Can we use all the data for a more through picture? - Can we rely on "noise" and "adversarial examples" to supply more nuance to our AI?	Breck et al. (2019) Goodfellow et al. (2015) Kaul (2018) Kulkarni et al. (2020)
Evaluation	- Evaluation as part of the AI process - Avoiding a built and done process - Creating an evaluation process for shared use - Incorporating AI assessment into both the construction and output process	Agarwal et al. (2018) Brennen (2020) Jacobs and Wallach (2019) Messick (1998) Zhang et al. (2018)

Technology moves fast, and with different advances come new challenges and the need for new ways to tackle assurance in AI. Though there are philosophical ramifications to tackling a topic like AI assurance, and though it is still challenging to come to a complete definition of "fairness," the suggestions in this chapter can help provide some initial steps for how validation can make a difference in how we tackle the problem of bias in AI. As a society, we have expectations of what is fair, even if we have a more difficult time agreeing on a definition. Due to the difficulty in describing fairness, it becomes harder to detect our biases in our day-to-day lives.

In many cases, the biases we hold, which we may not be able to quickly or easily identify, provide the opportunity for us to encode these biases into our algorithms unintentionally. There is also the more significant issue of assuming that our data, due to the availability and accessibility of large sample sizes, are accurate or free of bias. Additionally, we need to consider our use of training data and whether or not the sample we are using is representative, especially due to data integrity issues and biases in data collection (Mehrabi et al., 2019). This also applies to cases where data may not be available or accessible. We hope that despite these limitations, we have shown that there are ways of verifying and validating our algorithms, and even calibrating our models to ensure better, bias-free practices (Bird et al., 2016).

The validation methods explored in this chapter range from computationally simple to more complicated automated methods of reducing bias. The methods provided can also have reliable results across different model-

ing communities. It is important that data scientists and AI specialists focus on being transparent in reporting findings, having results that are validated across a wide range of thresholds (Malik, 2020), as well as allowing for information to be accessible in a simple way so that people can understand the "what" and "how" of the tools being used.

When we manage these tools wisely, we can gain new insights we might not have gained, thereby providing new knowledge about the consequences of our tests. "Thus, evidence of construct meaning is not only essential for evaluating the import of testing consequences, it also helps determine where to look for testing consequences" (Messick, 1998). AI assurance and the many components that makeup assurance are here to stay as our technologies and societal needs advance. The importance of understanding that our tools are not consequence-free is an essential first step in addressing bias in AI (Batarseh et al., 2021). The greatest challenge is for us to advance while also improving upon our systems to ensure that fairness and accuracy go hand in hand.

Our goal with this chapter is to provide a synthesis of some of the efforts made in AI assurance and to provide AI researchers, data scientists, policymakers, and practitioners with a greater understanding of the problem space. With that understanding, our hope is that transparency, dialogue, and providing both useful and fair solutions can be the way forward to a more bias-free world.

References

Abdollahi, B., Nasraoui, O., 2018. Transparency in fair machine learning: the case of explainable recommender systems. In: Zhou, J., Chen, F. (Eds.), Human and Machine Learning. In: Human–Computer Interaction Series. Springer International Publishing, Cham, pp. 21–35. http://link.springer.com/10.1007/978-3-319-90403-0_2.

Agarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D., 2018. Automated test generation to detect individual discrimination in AI models. arXiv:1809.03260 [cs]. http://arxiv.org/abs/1809.03260. arXiv:1809.03260.

Balci, O., 1997. Verification validation and accreditation of simulation models. In: Proceedings of the 29th Conference on Winter Simulation - WSC '97. Atlanta, Georgia, United States. ACM Press, pp. 135–141. http://portal.acm.org/citation.cfm?doid=268437.268462.

Batarseh, F., Freeman, L., Huang, C.H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8.

Batarseh, F.A., Gonzalez, A.J., 2018. Predicting failures in agile software development through data analytics. Software Quality Journal 26, 49–66. https://doi.org/10.1007/s11219-015-9285-3. http://link.springer.com/10.1007/s11219-015-9285-3.

Bird, S., Barocas, S., Crawford, K., Diaz, F., Wallach, H., 2016. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In: Proceedings of Workshop on Fairness, Accountability. New York.

Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H., 2020. Language (technology) is power: a critical survey of "bias". arXiv:2005.14050 [cs]. http://arxiv.org/abs/2005.14050. arXiv:2005.14050.

Box, G.E.P., 1976. Science and statistics. Journal of the American Statistical Association 71, 791–799. https://doi.org/10.1080/01621459.1976.10480949. http://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949.

Breck, E., Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M., 2019. Data validation for machine learning. In: Proceedings of SysML. https://mlsys.org/Conferences/2019/doc/2019/167.pdf.

Brennen, A., 2020. What do people really want when they say they want "explainable AI?" we asked 60 stakeholders. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA. ACM, pp. 1–7. https://dl.acm.org/doi/10.1145/3334480.3383047.

Byun, T., Rayadurgam, S., 2020. Manifold for machine learning assurance. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results. Seoul South Korea. ACM, pp. 97–100. https://dl.acm.org/doi/10.1145/3377816.3381734.

Calders, T., Verwer, S., 2010. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge 21, 277–292.

Carlsson, M., Rooth, D.O., 2008. Is It Your Foreign Name or Foreign Qualifications? An Experimental Study of Ethnic Discrimination in Hiring. Technical Report Discussion Paper No. 3810. Institute for the Study of Labor, Germany.

Chen, I., Johansson, F., Sontag, D., 2018. Why is my classifier discriminatory? Advances in Neural Information Processing Systems, 3539–3550.

Chouldechova, A., 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5, 153–163.

Chouldechova, A., Roth, A., 2018. The frontiers of fairness in machine learning. arXiv:1810.08810 [cs, stat]. http://arxiv.org/abs/1810.08810. arXiv:1810.08810.

De-Arteaga, M., Fogliato, R., Chouldechova, A., 2020. A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA. ACM, pp. 1–12. https://dl.acm.org/doi/10.1145/3313831.3376638.

Dobson, J.E., 2015. Can an algorithm be disturbed?: machine learning, intrinsic criticism, and the digital humanities. College Literature 42, 543–564. https://doi.org/10.1353/lit.2015.0037. https://muse.jhu.edu/content/crossref/journals/college_literature/v042/42.4.dobson.html.

Dressel, J., Farid, H., 2018. The accuracy, fairness, and limits of predicting recidivism. Science Advances 4. https://doi.org/10.1126/sciadv.aao5580.

Dwork, C., Ilvento, C., 2018. Fairness Under Composition. 20 pages. http://drops.dagstuhl.de/opus/volltexte/2018/10126/. artwork Size: 20 pages Medium: applica-

tion/pdf Publisher: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany Version Number: 1.0.

Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P., Lurie, N., 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology 9, 69–83. https://doi.org/10.1007/s10742-009-0047-1. http://link.springer.com/10.1007/s10742-009-0047-1.

Fogliato, R., G'Sell, M., Chouldechova, A., 2020. Fairness evaluation in presence of biased noisy labels. arXiv:2003.13808 [cs, stat]. http://arxiv.org/abs/2003.13808. arXiv:2003.13808.

Fujii, G., Hamada, K., Ishikawa, F., Masuda, S., Matsuya, M., Myojin, T., Nishi, Y., Ogawa, H., Toku, T., Tokumoto, S., Tsuchiya, K., Ujita, Y., 2020. Guidelines for quality assurance of machine learning-based artificial intelligence. International Journal of Software Engineering and Knowledge Engineering 30, 1589–1606. https://doi.org/10.1142/S0218194020400227. https://www.worldscientific.com/doi/abs/10.1142/S0218194020400227.

Go, W., Lee, D., 2018. Toward trustworthy deep learning in security. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto Canada. ACM, pp. 2219–2221. https://dl.acm.org/doi/10.1145/3243734.3278526.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. arXiv:1412.6572 [cs, stat]. http://arxiv.org/abs/1412.6572. arXiv:1412.6572.

Gore, R.J., Lynch, C.J., Kavak, H., 2017. Applying statistical debugging for enhanced trace validation of agent-based models. Simulation 93, 273–284. https://doi.org/10.1177/0037549716659707. http://journals.sagepub.com/doi/10.1177/0037549716659707.

Habli, I., Lawton, T., Porter, Z., 2020. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization 98, 251–256. https://doi.org/10.2471/BLT.19.237487. http://www.who.int/entity/bulletin/volumes/98/4/19-237487.pdf.

Hagras, H., 2018. Toward human-understandable, explainable AI. Computer 51, 28–36. https://doi.org/10.1109/MC.2018.3620965. https://ieeexplore.ieee.org/document/8481251/.

Jacobs, A., Wallach, H., 2019. Measurement and fairness. Computers & Society. https://arxiv.org/abs/1912.05511.

Kaul, S., 2018. Speed and accuracy are not enough! Trustworthy machine learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New Orleans LA USA. ACM, pp. 372–373. https://dl.acm.org/doi/10.1145/3278721.3278796.

Kulkarni, A., Chong, D., Batarseh, F.A., 2020. Foundations of data imbalance and solutions for a data democracy. In: Data Democracy. Elsevier, pp. 83–106. https://linkinghub.elsevier.com/retrieve/pii/B9780128183663000058.

de Laat, P.B., 2018. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philosophy & Technology 31, 525–541. https://doi.org/10.1007/s13347-017-0293-z. http://link.springer.com/10.1007/s13347-017-0293-z.

Lum, K., Isaac, W., 2016. To predict and serve? Significance 13, 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x. http://doi.wiley.com/10.1111/j.1740-9713.2016.00960.x.

Lynch, C.J., Diallo, S.Y., Kavak, H., Padilla, J.J., 2020. A content analysis-based approach to explore simulation verification and identify its current challenges. PLoS ONE 15, e0232929. https://doi.org/10.1371/journal.pone.0232929. https://dx.plos.org/10.1371/journal.pone.0232929.

Madaio, M.A., Stark, L., Wortman Vaughan, J., Wallach, H., 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA. ACM, pp. 1–14. https://dl.acm.org/doi/10.1145/3313831.3376445.

Madras, D., Pitassi, T., Zemel, R., 2018. Predict responsibly: improving fairness and accuracy by learning to defer. arXiv:1711.06664 [cs, stat]. http://arxiv.org/abs/1711.06664. arXiv:1711.06664.

Mahadevaiah, G., Rv, P., Bermejo, I., Jaffray, D., Dekker, A., Wee, L., 2020. Artificial intelligence-based clinical decision support in modern medical physics: selection, acceptance, commissioning, and quality assurance. Medical Physics 47, e228–e235. https://doi.org/10.1002/mp.13562.

Malik, M., 2020. A Hierarchy of Limitations in Machine Learning. Berkman Klein Center for Internet & Society at Harvard University. https://arxiv.org/pdf/2002.05193.pdf.

Malinas, G., The Hegeler Institute, 2001. Simpson's paradox: a logically benign, empirically treacherous hydra. The Monist 84, 265–283. https://doi.org/10.5840/monist200184217. https://academic.oup.com/monist/article-lookup/doi/10.5840/monist200184217.

McNamara, D., Ong, C.S., Williamson, R.C., 2017. Provably fair representations. arXiv:1710.04394 [cs]. http://arxiv.org/abs/1710.04394. arXiv:1710.04394.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2019. A survey on bias and fairness in machine learning. arXiv:1908.09635 [cs]. http://arxiv.org/abs/1908.09635. arXiv:1908.09635.

Messick, S., 1998. Test validity: a matter of consequence. Social Indicators Research 45, 35–44.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J.P., Rosenquist, J., 2011. Understanding the Demographics of Twitter Users, Spain.

Munoko, I., Brown-Liburd, H.L., Vasarhelyi, M., 2020. The ethical implications of using artificial intelligence in auditing. Journal of Business Ethics 167, 209–234. https://doi.org/10.1007/s10551-019-04407-1. http://link.springer.com/10.1007/s10551-019-04407-1.

Nabi, R., Shpitser, I., 2018. Fair inference on outcomes. arXiv:1705.10378 [stat]. http://arxiv.org/abs/1705.10378. arXiv:1705.10378.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S., 2020. Bias in data-driven artificial intelligence systems—an introductory survey. WIREs Data Mining and Knowledge

Discovery 10. https://doi.org/10.1002/widm.1356. https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1356.

Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E., 2019. Social data: biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2, 13. https://doi.org/10.3389/fdata.2019.00013. https://www.frontiersin.org/article/10.3389/fdata.2019.00013/full.

O'Neil, C., 2016. Weapons of Math Destruction: How Big Data Increases Inequality. Crown Books.

Osoba, O., Welser, W., 2017. The Risks of Artificial Intelligence to Security and the Future of Work. RAND Corporation. https://www.rand.org/pubs/perspectives/PE237.html.

Papakyriakopoulos, O., Serrano, J.C.M., Hegelich, S., 2020. Political communication on social media: a tale of hyperactive users and bias in recommender systems. Online Social Networks and Media 15, 100058. https://doi.org/10.1016/j.osnem.2019.100058. https://linkinghub.elsevier.com/retrieve/pii/S2468696419300886.

Pedreshi, D., Ruggieri, S., Turini, F., 2008. Discrimination-aware data mining. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08. Las Vegas, Nevada, USA. ACM Press, p. 560. http://dl.acm.org/citation.cfm?doid=1401890.1401959.

Rodrigues, R., 2020. Legal and human rights issues of AI: gaps, challenges and vulnerabilities. Journal of Responsible Technology 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005. https://linkinghub.elsevier.com/retrieve/pii/S2666659620300056.

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., Kalai, A.T., 2019. What's in a name? Reducing bias in bios without access to protected attributes. arXiv:1904.05233 [cs, stat]. http://arxiv.org/abs/1904.05233. arXiv:1904.05233.

Shore, C., Wright, S., 2015. Audit culture revisited: rankings, ratings, and the reassembling of society. Current Anthropology 56, 421–444. https://doi.org/10.1086/681534. https://www.journals.uchicago.edu/doi/10.1086/681534.

Srivastava, B., Rossi, F., 2019. Towards composable bias rating of AI services. arXiv:1808.00089 [cs]. http://arxiv.org/abs/1808.00089. arXiv:1808.00089.

Strathern, M., 1997. Improving ratings: audit in the British university system. European Review 5.

Varshney, K.R., 2019. Trustworthy machine learning and artificial intelligence. XRDS: Crossroads, The ACM Magazine for Students 25, 26–29. https://doi.org/10.1145/3313109. https://dl.acm.org/doi/10.1145/3313109.

Wing, J.M., 2020. Trustworthy AI. arXiv:2002.06276 [cs]. http://arxiv.org/abs/2002.06276. arXiv:2002.06276.

Yang, K., Stoyanovich, J., 2017. Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management. Chicago IL USA. ACM, pp. 1–6. https://dl.acm.org/doi/10.1145/3085504.3085526.

Zemel, R., Yu, W., Swersky, K., Pitassi, T., Dwork, C., 2013. Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning. http://proceedings.mlr.press/v28/zemel13.html.

Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning. arXiv:1801.07593 [cs]. http://arxiv.org/abs/1801.07593. arXiv:1801.07593.

Ören, T.I., 1987. Quality assurance paradigms for artificial intelligence in modelling and simulation. Simulation 48, 149–151. https://doi.org/10.1177/003754978704800406. http://journals.sagepub.com/doi/10.1177/003754978704800406.

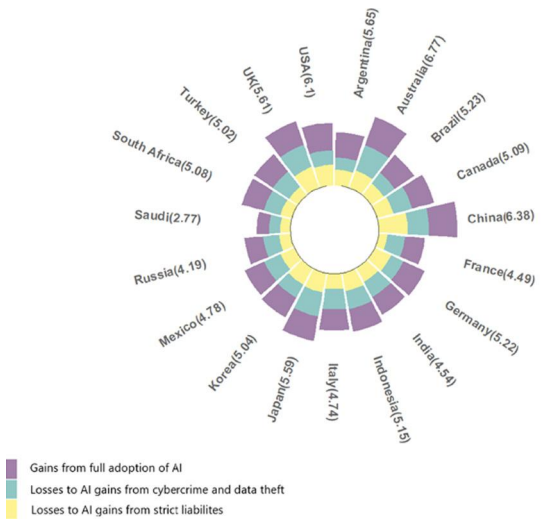
This page intentionally left blank

5

An evaluation of the potential global impacts of AI assurance

Sindhu Bharathi, Badri Narayanan, Sumathi Chakravarthy, and Shounkie Nawani
Infinite Sum Modelling LLC, Seattle, WA, United States

Graphical abstract



Abstract

As much as Artificial Intelligence is estimated to transform humanity, ethics of these systems are questioned, and strict regulations are univocally called upon.This chapter aims to analyze the ethical frameworks adopted by countries across the world and quantify using a CGE model, the benefits from full adoption of AI across the world, the gains that each country could make from erecting regulatory frameworks that govern ethical usage of AI, and the losses from imposition of such strict liabilities. Results reveal that there would be enormous gains from complete adoption of AI across the world. Australia, China, and USA would

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00017-2

gain the most. Countries that experience data theft and cybersecurity crimes, for example, Argentina, Brazil, USA, and Germany, experience higher losses. As against the conventional argument that automation could lower employment rates, the model estimates a rise in employment of skilled labor.

Keywords

Ethical AI, AI economics, policy shocks, liability

Highlights

- Under complete adoption of Artificial Intelligence, the Australia, China and USA would gain the highest and their GDP is estimated to increase by about 6.77
- Where cybercrime and data theft is a problem, to benefit and leverage the fullest potential of AI, countries require legal frameworks to govern the development and usage of AI systems.
- Argentina, France, Australia, Turkey, and South Africa experience higher losses due to imposition of strict liabilities, as it slows down adoption of the technology and further innovation. In certain countries, such as Australia, Japan, Korea, India, Canada, Mexico, France, Russia, Turkey, and South Africa, losses from strict liabilities outweigh the benefits, and so it becomes essential for regulatory institutions to erect a harmonized legal framework and ensure that such frameworks do not end up being adoption barriers.

5.1 Introduction

Artificial Intelligence is broadly perceived as computational systems that display higher levels of intelligence, including "narrow AI," which demonstrates and excels at the automation of certain specified tasks, and "general AI," which produces an intelligent agent with higher ability to learn and reason (Davidson, 2019).

AI is gaining momentum with widespread adoption across the globe and is anticipated to revolutionize and transform the way we live and work in unprecedented ways. More evident than ever before, the greater proliferation of AI to virtually disrupt mainstream business processes have made it a mandate for organizations to fully harness the potential of AI and other disruptive technologies to stand apart in an extremely competitive world.

The public budgetary allocation varies across different countries, from $500 million in countries such as Japan, Korea, and the UK to around USD 1 million in countries such as Australia, Estonia, Lithuania, Portugal, and Greece (OECD, 2021).

The rapid technological advancements and scientific breakthroughs in decision-making capabilities have brought AI to the forefront of digitization and the fourth industrial revolution. It's scope in transforming the business activities and industrial operations to greater heights, and in gravitating the future to enormously higher degrees of automation are paramount. From increasing the efficiency of farming through crop and soil monitoring, decoding crimes through predictive analytics, to diagnosing diseases and supporting minimally invasive surgeries, AI-powered tools and solutions are finding adoption in improving our lives in myriad ways by performing operations that were once considered impossible.

PWC (2018) predicts that adoption and application of AI would bring about an increase in the global GDP by 14% by 2030, adding $15.7 trillion to the economy. Bughin et al. (2018) has estimated an increase in the global economic output of $13 trillion, and 1.2% increase in global GDP every year. It has predicted that the adoption may follow a S-curve, meaning that though the initial investments may be slow, its adoption will increase with an accelerated investment at later stages in the attempt to bag competitive advantages. It also advocates that those early adopters and frontrunners would make disproportionate gains and additional economic benefits between 20–25% compared to the rest. Several studies have predicted the enormous scope and scale of the impact of AI on businesses, economy, and society and have advocated their concerns that ex-ante regulations may hinder the broad-scale adoption of AI. Countries across the globe have channelized their efforts and resources in harnessing the fullest potential of AI to gain technological leadership on one hand and on the other, are strategizing action plans, innovative policies, and regulatory measures to mitigate the risks concerning irresponsible use of AI in certain crucial operations.

Though AI has the power to promote inclusive growth, it presents new challenges to formal governance procedures and incumbent legislative

frameworks particularly in terms of ensuring transparency, privacy, and security to everybody involved. The increasing adoption in performing crucial operations in healthcare, national security, etc., pose potential risks, and countries are in the process of putting forth regulatory measures to confront such risks and ensure a safe democratization of AI. The intrusion of such technologies into the privacy of our lives, and the higher chances of it being mishandled by criminals to execute illegitimate activities have also raised concerns.

Methods to build trust and understanding become even more significant as these technologies are increasingly used to design mainstream solutions that demand automated decision-making and problem solving, rather than merely being perceived as supportive tools and solutions. In light of the changing economic and social landscape, there is a significant need to draw meaningful policy suggestions from research in this domain and propose policies that shall go hand-in-hand with the existing ones spearheaded towards achieving an accelerated penetration of such technologies.

The implementation of any new path-breaking technology requires development of well-functioning policies, legal frameworks, and sound governance to ascertain that the development, deployment, and diffusion of such technologies has a positive impact on people and countries all around the world.

This study aims to understand the best practices and ethical frameworks adopted by countries across the world to streamline AI adoption and to leverage the benefits such technologies are capable of offering, to the fullest possible extent. We use a computable general equilibrium (CGE) model, the Global trade analysis project (GTAP), to estimate the possible gains from full adoption of AI technology across each country/region in the world. Because the model is capable of capturing the inter-sectoral linkages between different regions, it becomes possible to carry out analysis at a granular level.

The model has 141 regions/countries and 65 sectors referenced to the year 2014. We aggregated the GTAP database to finally cover 26 regions and 24 regions, with a focus on G-20 countries. The impact of AI across the regions at the sector-level are analyzed in three scenarios: (1) first considers

complete adoption of AI across all the countries to quantify the economic benefits of AI athwart all industry sectors and its positive impact on variables such as output (GDP), productivity, trade, investments, innovation, consumer welfare, and increase in worker wages; (2) second considers the gains that each country can make from adopting ethical frameworks and standards to curb illegitimate usage of such technologies and to regulate their adoption; and (3) third scenario estimates the losses that each country/region would endure due to the imposition of strict liabilities and restrictive legal norms, which in turn could curb countries from leveraging the true potential of AI.

Results of the study reveal that there will be a gain in GDP of all the regions when full adoption is assumed in scenario 1. Australia, China, and USA would gain the most to about 6.77%, 6.38%, and 6.1% of GDP, respectively. In scenario 2, the study estimates a loss in GDP to all countries when cybercrimes and data thefts are left unprotected. Argentina, Brazil, USA, and Germany experience higher losses as cybercrime rates as a percentage of their GDP is higher in these countries. Such losses could be mitigated by imposing and strategizing legal frameworks to govern the development and usage of AI systems. In scenario 3, when the model is shocked to estimate the losses due to imposition of strict liabilities and ex-ante regulations, Argentina, France, Australia, Turkey, and South Africa experience higher losses. In some countries, the loss in GDP in scenario 3 is higher than in scenario 2, meaning the losses that incur from restrictive legal practices is higher than the losses from cybercrime attacks. The rate of employment increases in scenario 1 against the typical argument and concern raised by most experts stating that automation could lower employment rates. The rise in employment in the model is due to the employment of more skilled labor.

The rest of the chapter is organized as follows: Section 5.2 covers the literature review and details the best practices adopted by each country to regulate and monitor AI related applications. Section 5.3 covers the methodology and details the intricate details of the GTAP model, scenarios, and assumptions we have made to take forward the analysis. Section 5.4 presents the results of three different scenarios and estimates of the change in GDP,

exports, imports, employment and sectoral output for each country due to the adoption of AI as well as due to ethical frameworks, and regulatory norms, whereas Section 5.5 concludes and presents final remarks.

5.2 Literature review

This chapter leverages descriptive and qualitative analysis to understand the ethical impacts of AI, and to propose a strategic and policy framework to deploy AI. The following provides an outline of the legislative norms and relevant ethical principles that are adopted by countries or regional institutions to regulate the development and usage of AI.

The European Commission in April 2021 published its first legal framework, following the white paper that was released in February 2020 to regulate and put forth a comprehensive package to ensure trustworthy AI. They follow a risk-based approach by classifying the applications under unacceptable, high-risk, limited risk, and minimal risk. Those applications that include systems that are a clear threat to safety and to the rights of the citizens would be brought under unacceptable risk: AI systems, including critical infrastructure, educational training, safety components of products, employment, workers management, and access to self-employment, essential private and public services, migration, asylum and border control management. Those with specific transparency commitment, such as chatbots are brought under limited risk and those that allow free usage of applications, such as video games, etc. are categorized under no risk or minimal risk category. By doing so, the commission intends to address risks with a flexible set of rules based on specific applications or usage of AI (European Commission, 2020).

USA has the highest number of AI policy institutions and initiatives, followed by Austria and the United Kingdom. The policy landscape is rather decentralized, and the country's policy emphasizes collaboration between federal agencies, academia, the private sector, and non-profit organizations to bolster an innovation ecosystem (OECD, 2021). The Department of Defence (DoD), USA, adopted 5 principles of AI based on recommendation of Defence Innovation Board (DIB): responsible, equitable, traceable, reliable, and governable (U.S. Department of Defense, 2020).

Indonesia's economy is expected to shift into a higher gear with AI, with an expected US$366 billion added to the country's gross home product by 2030 (Hunt, 2020). Though AI adoption is still in its infancy in Southeast Asia, more than 70% of AI users, suppliers, and investors regard it as critical to the region's future. Recently, the Indonesian government introduced a national strategy that will guide the country in developing AI between 2020 and 2045, focusing on education and research, health services, food security, mobility, smart cities, and public sector reform. In terms of progress in AI, a 2018 International Data Corporation survey found that Indonesian companies had the highest rates of AI adoption in Southeast Asia, with 24.6% of organizations integrating the technology into their operations. Indonesia has joined the ranks of countries such as Singapore, South Korea, and Canada in developing national AI plans. While many governments' AI policies focus on economic development, Indonesia's strategy stands out for its focus on using AI to solve specific problems. For example, it wants to employ AI to assist solve its problems with child malnutrition and to digitize government services.

The United Kingdom has played an essential part in the history and development of AI. In addition to the UK having been extensively involved in AI development from the beginning, it also contributed to the industry's first AI Winter (gov. uk). AI's promises resulted in a significant reduction in government, research, and university funding. The study had a gloomy approach to AI and was harshly critical of several key elements of research in the field. With the return of interest and investment in AI, the UK has responded by making significant investments in the area, demonstrating its strength in the sector. According to a McKinsey Global Institute study, the United Kingdom has one of the best AI strategies in the world, with significant government support for AI, research activity in the area, VC financing for AI companies, and business AI activity and adoption (McKinsey Global Institute, 2019). The UK formed an All-Party Parliamentary Group on AI in 2017 to discuss ethical problems, industry standards, regulatory alternatives, and societal effect of AI in Parliament (Cain et al., 2020).

The Chinese State Council in 2017 proposed the New generation AI development plan (NGADP), for adopting AI (AI) ethics principles. The "Bei-

jing AI Principles" given by the Beijing Academy of AI (BAAI), which is supported by the Chinese Ministry of Science and Technology and the Beijing municipal government states the guidelines for research and development in AI, including major rights to be respected on human privacy, dignity, freedom, autonomy, etc. (Arcesati, 2021).

In 2018 AI standardization white paper issued by the Chinese Electronics Standards Institute (CESI) recommended three significant ethical considerations for AI: "human interest," "liability," and "consistency of rights and responsibilities," which extensively discusses the safety measures, ethical, and privacy issues and highlights the government's wish to use technical standardization as a tool in domestic and global AI governance (Gal, 2020).

The Japanese Society for AI (JSAI) has formalized the ethical guidelines for its members, which serves as a moral foundation for JSAI members to use AI effectively, while understanding the social and ethical responsibilities (ELSI, 2021). These include points such as i) contribution to humanity (human rights), ii) abidance of laws and regulations (IPR and R&D), iii) equality, and iv) security. Japan also follows the OECD AI principles and aims at establishing an AI economic society, where all use AI and data users actively participate in social and economic activities.

Russian President Vladimir Putin called on the international communities to help foster AI and restrict its use for the benefit of humanity; he also asked UN members to seek AI regulations, which support military and technological security, law, and morality (Bendett, 2020). In October 2019, Russian Government released a national AI strategy on the development of AI in the Russian Federation, which focuses on the future goals of developing AI and forming a regulatory system that guarantees public safety and also helps in stimulating the development of AI (Bendett, 2020).

The Canadian government in 2018 started an ethical analysis in AI, by examining the data storage at the Department of National Defence. In January 2020, the Office of the Privacy Commissioner (OPC) launched a consultation on the regulation of AI and enhanced some policy measures, such as human rights of privacy, transparency, and other social beneficiary topics, which were earlier looked after by the Personal information protection and electronic data act (PIPEDA) that governs the data privacy in Canada.

In January 2019, Singapore with the help of the Personal Data Protection Commission (PDPC) released the Model AI governance framework, which focuses primarily on four broader areas: internal governance, decision-making models, operations management, and customer relationship management (PDPC, 2019). The Cybersecurity strategy of Singapore along with digital security infrastructure helps in maintaining the Cyber Watch Centre's operational excellence and also helps in timely detection of and response to a cyber incident.[1]

AI has the potential to add $957 billion, which is 15% of India's GDP in 2035. In 2018–19, the finance ministry of India proposed a spending of ₹7000 crore ($1 billion) for the next five years, which is very low when compared to the investment made by other developing countries, including China and USA (Menon and Roy, 2021). If India is to catch up with China and other nations in the growing area of AI, it will need to invest heavily in creating all the essential enabling technologies and eco-systems, as well as a framework for AI ethics and standards. AI has recently been used on a modest but successful scale in a variety of industries, ranging from robotic concierges in hotels to automated entertainment or cell phones. AI has reshaped a variety of sectors (Global Legal Insights, 2021). The NITI Aayog has suggested establishing an oversight organization to establish standards, rules, and benchmarks for the usage of AI across industries, which will be required for government procurement. Field experts from computer science, AI, legal experts, sector specialists, and representatives from civil society, humanities, and social science are likely to make up the body. The oversight group, according to the proposed paper, must serve an enabling role in the broad fields of AI research, technical, legal, policy, and societal concerns. Additionally, it should clarify responsible behavior through design structures, standards, and guidelines, as well as give access to responsible AI tools and approaches (Mondaq, 2020).

Hwang and Park (2019) examined various AI charter of ethics (AICE) in the Republic of Korea in suggesting response to threats from AI. AI threats are classified into three categories: Firstly, AI's value judgment, malicious

[1] https://indiaai.gov.in/country/singapore?standard=interoperability.

use of AI, and human alienation. Secondly, Korea's seven AICEs with the objective to create, develop, and utilize AI in a manner ensuring human safety and improves well-being by accurately identifying the positive and negative impacts that may arise from AI, and Korea has got seven documents classified as AICEs.

Seven Al charters of ethics in South Korea:

1. Draft of the robot ethics charter (DREC) (Ministry of Commerce Industry and Energy, 2007 March) identifies human-centered ethical codes for existence of humans and robots.
2. Kalao algorithm ethics charter (KAEC) applies social ethics to all efforts related to algorithms to benefit and bring happiness for humankind.
3. Ethical guidelines - for intelligence information society (EGIIS) aims to realize the value of sustainable symbiosis to move towards safe and reliable intelligent information.
4. Intelligent government ethics guidelines: for utilizing AI (IGEG) aims to respond to problems that are caused by AI, using government services according to the basic plans of intelligent government announced in March 2007.
5. Charter of AI ethics (CAIE) aims to find adverse effects of AI and find ways to respond to them.
6. Principles for user oriented intelligence society (PUOES) suggest public rules for a safe intelligent information society protected from the risks that are caused by adoption of new technology.
7. Ethical guidelines for self-driving cars aims to improve human safety and welfare, to ensure safe and convenient freedom on right of mobility, to consider human life first before animals lives or property damage, and to minimize personal and social loss from accidents.

AI in Africa's health care can improve various aspects of health care. It can reduce annual expenditure, allow early detection of disease, provide around-the-clock monitoring of chronic disorders and help limit the exposure of healthcare professionals in a contagious environment. In Africa, the main focus of AI is in healthcare industry to eliminate inefficiencies in misdiagnosis, shortage of healthcare workers, and wait for the recovery time. However, it is important to safeguard against issues such as privacy

breaches, lack of personalized care, and accessibility and this study is to suggest policy makers to strike a balance between allowing innovation and protecting data (Observation Research Foundation, 2019).

The main objective of the Australia's AI ethics framework is to perceive key governance canons and measures that can be used to accomplish the best conceivable results from AI, while keeping the well-being of Australians as the top precedence and designing safer and reliable outcomes. These principles mainly focus on human, societal, and environmental well-being, privacy protection and security, reliability and safety, transparency, and explain ability and contestability and accountability. It further ensures that AI systems should respect and uphold privacy rights and data protection and safeguard the security of data. Adequate access to information on the AI algorithm and inferences drawn is mandatory to ensure contestability and design an effective system of oversight so as to make appropriate use of human judgment. Responsibility and accountability for AI systems and their outcomes both before and after design, development, and operation is mandated (Commonwealth Scientific and Industrial Research Organisation, 2019).

5.3 Methodology & modeling

In this study, we have used a multi-sector, multi-regional computable general equilibrium (CGE) model to analyze and estimate the gains to each country/region and the losses that incur from strict liability imposed by legal regulatory norms to govern the usage and adoption of AI. We used an extension of the standard GTAP framework designed and developed by the Center for Global Trade Analysis to be used by researchers and economists to study the impact of trade policies and frameworks. The fact that the model is so widely used by international organizations, such as the United Nations Conference on Trade and Development (UNCTAD), World Bank, World Trade Organization (WTO), and Organization for Economic Co-operation and Development (OECD) speaks volumes about its reliability and effectiveness.

The widely used GTAP modeling approach was designed and developed by Professor Hertel, head of the Global trade analysis project from Pur-

due University. The model generates impact results for national account aggregates, industry output and prices, factor inputs and prices, and trade flows. For a technical description of the GTAP model, see Hertel (1997); for a discussion of the degree of confidence in CGE estimates, see Hertel et al. (2004).

The GTAP model designed by Hertel is executed and implemented in real-time using GEMPACK (general equilibrium modeling package), a package that has a suite of economic modeling programs and software designed, developed, and provided by Centre of Policy Studies (CoPS), Victoria University. The standard case in a static CGEs is that savings determine the investment demand, but that the capital stock is fixed, and thus not linked to changes in investment.

The GTAP model is characterized by perfect competition, constant returns to scale and Armington elasticities. Such a multiregional, multisector, computable general equilibrium (CGE) model can capture the macroeconomic aspects, the supply-chain effects, factor-use effects of various commodities, economy-wide equilibrium constraints, apart from the complete linkages between different sectors and countries. The model assumes inter-sectoral substitution and so is also able to capture the potential substitution of one sector by another, which is a significant aspect.

We use the latest and publicly available data from the GTAP 10 database (Aguiar et al., 2019), which contains global trade data for the years 2004, 2007, 2011, and 2014, with input-output tables and data on the current applied levels of trade protection.

The GTAP 10 database covers 141 regions/countries, and 65 sectors. For the convenience of analysis, the 141 regions are aggregated into 26 countries/regions as follows: Australia, China, Japan, Korea, Indonesia, India, Canada, USA, Mexico, Argentina, Brazil, France, Germany, Italy, the UK, Russia, Saudi Arabia, Turkey, South Africa, Rest of EU, Rest of Oceania, Rest of Asia, Rest of America, Rest of World, Middle East and North Africa, and Sub-Saharan Africa.

The 65 sectors are aggregated into 24 sectors as below: Agriculture, Extraction, Consumer Packaged Goods, Light Manufacturing, Other Manufacturing, Chemicals, Pharmaceutical and Medical Products, Basic Metals,

Automobiles, Computer Manufacturing, Electrical Equipment, Machinery, Services, Trade, Travel, Transport & Logistics, Tele-Communication, Banking, Insurance, Business Services, Media & Entertainment, Public and Social Services, Education, Health and Social services.

The 2014 data that is available in the GTAP 10.0 database is scaled to 2017 with data provided from World Bank and International Monetary Fund (IMF). The GTAP model effectively captures the direct linkages and indirect interactions in the economy. The model is widely preferred for policy analysis owing to its unique capability to effectively model supply-chain effects, macro-economic aspects, economy-wide equilibrium constraints, linkages between different sectors and countries, and the factor-use effects of various commodities to predict economic variables, such as GDP, trade balances, investments, innovation, consumer welfare, productivity, employment, and wages.

GTAP's ability to capture both sectoral and regional linkages help understand the impact of AI adoption of each country, not just pertaining to their individual socioeconomic realities, but also considering the socioeconomic linkages across other regions.

5.3.1 Scenario 1: full adoption of AI across all regions

Chui et al. (2018) has estimated that AI has the potential to create between USD 3.5 trillion and USD 5.8 trillion across nineteen industries. The study predicts that AI could add about $13 trillion to the global output by 2030. The paper analyzes more than 400 use cases across 19 industries and nine business functions. In the study, a "use case" is referred to as a targeted application of digital technologies to a specific business challenge, with a measurable outcome. Based on this definition, each use case was further bucketed on what analytical techniques could be used, traditional versus deeper machine learning forms. The use of these techniques varied based on the industry and function in which the use case was used, and a range was created to capture this variation in use of analytical techniques. And finally, these AI techniques determined the performance/productivity improvement observed by each industry.

Table 5.1 Estimated AI gains across different sectors.

Sector	AI Impact in % (of Industry revenues)	Aggregate dollar impact ($ Trillion)
Retail	3.2–5.7	0.4–0.8
Transport & Logistics	4.9–6.4	0.4–0.5
Travel	7.2–11.6	0.3–0.5
Consumer packaged goods	2.5–4.9	0.2–0.5
Public & Social Sector	1.1–1.4	0.3–0.4
Automotive & Assembly	2.6–4.0	0.3–0.4
Healthcare systems & services	2.9–3.7	0.2–0.3
Banking	2.5–5.2	0.2–0.3
Advanced electronics and semiconductors	3.3–5.3	0.2–0.3
Basic materials	1.6–3.1	0.2–0.3
High tech	5.7–10.2	0.2–0.3
Oil and Gas	1.8–1.9	0.2–0.2
Insurance	3.2–7.1	0.1–0.3
Agriculture	2.4–3.7	0.1–0.2
Chemicals	1.0–2.3	0.1–0.2
Media and Entertainment	2.9–6.9	0.1–0.2
Telecommunications	2.9–6.3	0.1–0.2
Pharmaceuticals and Medical Products	4.2–6.1	0.1–0.1
Aerospace and Defense	1.8–3.2	<0.1T

Source: McKinsey Global Institute study.

Table 5.1 summarizes of sector-level percent gain estimates due to the adoption of AI.

Scenario 1 in our study quantifies the impact of full adoption of AI technology across all countries based on the sectoral total factor productivity (TFP) shocks derived from the above-mentioned McKinsey study. We shocked our model with the average of the limits mentioned in the table above.

5.3.2 Scenario 2: estimation of gains from AI ethical frameworks across all regions

To estimate the gains that each region could derive from adoption of ethical frameworks, we first identify the losses that each region could endure

if there were no such regulatory policies. Morgan (2020) has estimated that cybercrimes are a major threat to the economy as they are capable of inflicting global economic damage of $6 Trillion USD in 2021 and could escalate to $10.5 Trillion USD by 2025. These costs include stolen money, data destruction, theft of personal data, theft of financial data, attacks that distort the normal functioning of the business, reputational harm, restoration of the system after being hacked or damaged, theft of intellectual property, etc.

These costs are so high that when measured as a country, this ranks—after the USA and China—as the third largest world economy. Given the high magnitude of cybercrime losses, we estimate the economic losses to AI's estimated potential gains due to cybercrime attacks. Center for Strategic and International Studies (2014) have estimated the economic impact of cybercrime attacks by measuring cybercrime losses as a percentage of each country's GDP. Swiatkowsa (2020) have estimated the regional distribution of cybercrime by estimating losses as a percentage of each regional GDP.

Dean et al. (2012) has estimated the internet economic activity of 2016 as a percentage of GDP for various G20 countries, which altogether were estimated to 4.2 trillion. In some countries such as the UK, the internet economy is as high as 12.4% of its GDP. We use the data of cybercrime losses and internet gains from the above-mentioned studies to estimate the percentage of losses to internet gains from data and cybercrime attacks. We then calculate the losses that could incur to AI gains using the estimated losses from above, and the GDP gains to each country/region from the estimates in scenario 1. Using the factor input change per region from the results, we derive the sectoral shocks or the sectoral losses to AI from data breaches and cybercrimes (see Table 5.2).

$$\begin{aligned} &\textit{AI losses from cybercrimes} \\ &\quad = \textit{Loss from cybercrimes} * \textit{economic benefits from AI} \\ &\textit{AI losses in terms of GDP} \\ &\quad = \textit{AI losses from cybercrimes} * (-\textit{Estimated GDP value}) \end{aligned}$$

Table 5.2 Shock to calculate estimated AI gains from a harmonized regulatory framework.

Countries	Loss from Cybercrime attacks (%)	Economic benefits (%)	Loss to Internet gains (%)	GDP estimate from Scenario 1 (%)	Loss to AI gains (%)	Factor input change (%)
Australia	0.08	3.7	0.022	6.77	−0.146	−0.22
Rest of Oceania	0.09	3.7	0.024	5.1	−0.124	−0.12
China	0.63	6.9	0.091	6.38	−0.583	−0.19
Rest of Asia	0.53	5.6	0.095	5.16	−0.488	−0.39
Japan	0.02	5.6	0.004	5.59	−0.02	−0.04
Korea	0.71	8	0.089	5.04	−0.447	−0.34
Indonesia	0.53	10	0.053	5.15	−0.273	−0.32
India	0.21	5.6	0.038	4.54	−0.17	−0.2
Canada	0.17	3.6	0.047	5.09	−0.24	−0.15
USA	0.64	5.4	0.119	6.1	−0.723	−0.45
Mexico	0.17	4.2	0.04	4.78	−0.193	−0.16
Rest of America	0.42	4.2	0.1	5.04	−0.504	−0.4
Argentina	0.42	3.3	0.127	5.65	−0.719	−0.49
Brazil	0.32	2.4	0.133	5.23	−0.697	−0.22
Rest of EU	0.41	5.7	0.072	5.49	−0.395	−0.32
France	0.11	3.4	0.032	4.49	−0.145	−0.17
Germany	0.41	4	0.103	5.22	−0.535	−0.39
Italy	0.04	3.5	0.011	4.74	−0.054	−0.08
UK	0.16	12.4	0.013	5.61	−0.072	−0.12
Rest of the World	0.84	5.7	0.147	5.04	−0.743	−0.54
Russia	0.1	2.8	0.036	4.19	−0.15	−0.17
MENA	0.11	3.8	0.029	4.08	−0.118	−0.15
Saudi	0.17	3.8	0.045	2.77	−0.124	−0.18
Turkey	0.07	2.3	0.03	5.02	−0.153	−0.13
Sub Saharan Africa (SSA)	0.14	2.5	0.056	5.78	−0.324	−0.27
South Africa	0.14	2.5	0.056	5.08	−0.284	−0.25

5.3.3 Estimation of loss due to strict liabilities across all regions

Evas (2020) reveals that the existing regulatory framework in EU that imposes strict liabilities on applications of AI and robotics could cost 0.04% to

GDP of the region or in other words, if there is a better harmonized regulatory policy, the region could gain 0.04% of GDP.

Lee-Makiyama and Narayanan (2020) examined the cost of ex-ante regulations and strict liabilities by calculating the economic losses that could happen due to a shift from ex-post regulations to ex-ante regulations in online services. The study estimated that such costs could be as high as 0.56% of EU's GDP, and among countries, France could experience a loss of 0.89% in its GDP, Italy about 0.43%, the UK and Germany could experience a loss of about 0.41% of GDP.

To arrive at a realistic estimate of the losses to AI gains due to imposition of strict liability frameworks, we take an average of the loss estimates from the afore-mentioned papers. The sector-wise losses are estimated using the same procedure adopted in scenario 2 to estimate the shocks (see Table 5.3).

$$\begin{aligned} &\textit{AI losses from strict liabilities} \\ &\quad = \textit{Loss from strict liabilities} * \textit{Economic benefits from AI} \\ &\textit{AI losses from strict liabilities in terms of GDP} \\ &\quad = \textit{AI losses from strict liabilities} * (-\textit{Estimated GDP value}) \end{aligned}$$

5.4 Results and analysis

All the simulations/scenarios are performed with year 2017 as the baseline, and in the first scenario we modeled the benefits that each country/region could gain from full adoption of AI technologies. In the second scenario, the model is shocked with estimated losses in AI gains from data breach and cybercrime attacks to estimate the possible gains from erecting a harmonized regulatory framework to govern AI applications, and in the last scenario the losses from imposition of strict liabilities and ex-ante regulations are considered.

5.4.1 Impact of policy shocks on GDP of countries/regions

An analysis of the impact of full adoption of AI across all countries/regions in scenario 1 reveals that all countries experience a gain in GDP. The model estimates that Australia will experience the maximum gains of 6.77% of its

Table 5.3 Shocks to calculate estimated AI losses due to strict liabilities and ex-ante regulations.

Countries	Loss from strict liabilities (%)	Economic benefits (%)	Loss to Internet gains (%)	GDP estimate (%)	Loss to AI gains (%)	Factor input change (%)
Australia	0.30	3.7	−0.081	6.77	−0.549	−0.31
Rest of Oceania	0.30	3.7	−0.081	5.1	−0.414	−0.32
China	0.30	6.9	−0.043	6.38	−0.277	−0.16
Rest of Asia	0.30	5.6	−0.054	5.16	−0.276	−0.23
Japan	0.30	5.6	−0.054	5.59	−0.299	−0.22
Korea	0.30	8	−0.038	5.04	−0.189	−0.17
Indonesia	0.30	10	−0.030	5.15	−0.155	−0.2
India	0.30	5.6	−0.054	4.54	−0.243	−0.22
Canada	0.30	3.6	−0.083	5.09	−0.424	0.09
USA	0.30	5.4	−0.056	6.1	−0.339	−0.23
Mexico	0.30	4.2	−0.071	4.78	−0.341	−0.27
Rest of America	0.30	4.2	−0.071	5.04	−0.360	−0.28
Argentina	0.30	3.3	−0.091	5.65	−0.514	−0.34
Brazil	0.30	2.4	−0.125	5.23	−0.654	−0.07
Rest of EU	0.30	5.7	−0.053	5.49	−0.289	−0.23
France	0.47	3.4	−0.138	4.49	−0.621	−0.48
Germany	0.23	4	−0.058	5.22	−0.300	−0.23
Italy	0.24	3.5	−0.069	4.74	−0.325	−0.27
UK	0.23	12.4	−0.019	5.61	−0.104	−0.11
Rest of the world	0.30	5.7	−0.053	5.04	−0.265	−0.21
Russia	0.30	2.8	−0.107	4.19	−0.449	−0.36
MENA	0.30	3.8	−0.079	4.08	−0.322	−0.27
Saudi	0.30	3.8	−0.079	2.77	−0.219	−0.21
Turkey	0.30	2.3	−0.130	5.02	−0.655	−0.51
Sub-Saharan Africa	0.30	2.5	−0.120	5.78	−0.694	−0.42
South Africa	0.30	2.5	−0.120	5.08	−0.610	−0.45

GDP, China will gain by 6.38%, and USA by 6.10%. Argentina is expected to gain by 5.65% and the UK will gain by 5.61%.

In scenario 2, when shocked with estimated loss from AI gains due to data theft and cybercrime with data on economic losses from Morgan

Table 5.4 Impact of policy shocks on GDP of countries/regions.

Countries	Base Value (in millions USD)	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)
Australia	1,330,295	6.77	−0.78	−2.58
Rest of Oceania	259,429	5.10	−0.50	−1.72
China	12,143,477	6.38	−1.60	0.09
Rest of Asia	3,477,350	5.16	−2.00	−1.18
Japan	4,859,953	5.59	−0.07	−1.21
Korea	1,530,731	5.04	−1.70	−0.78
Indonesia	1,015,416	5.15	−1.05	−1.07
India	2,652,244	4.54	−0.64	−1.02
Canada	1,646,867	5.09	−0.78	−1.32
USA	19,485,402	6.10	−2.85	−1.45
Mexico	1,157,740	4.78	−0.77	−1.33
Rest of America	1,943,690	5.04	−1.95	−1.49
Argentina	642,691	5.65	−2.88	−2.13
Brazil	2,053,594	5.23	−2.17	−1.90
Rest of EU	14,323,426	5.49	−1.69	−1.30
France	26,792	4.49	−0.54	−2.50
Germany	255,756	5.22	−2.14	−1.20
Italy	141,510	4.74	−0.19	−1.31
UK	2,666,217	5.61	−0.34	−0.56
Rest of the world	1,765,886	5.04	−2.82	−1.04
Russia	1,578,624	4.19	−0.55	−1.65
MENA	2,515,801	4.08	−0.43	−1.18
Saudi	688,588	2.77	−0.26	−0.58
Turkey	852,674	5.02	−0.59	−2.66
SSA	1,343,885	5.78	−1.29	−2.72
South Africa	349,552	5.08	−1.17	−2.48

(2020), all countries experience a loss in GDP. Argentina experiences a loss of 2.88% and USA a loss of 2.85% in GDP. Brazil is expected to experience a decline of 2.17% of GDP, and Germany a loss of 2.14% in GDP. If regulatory policies and legal frameworks are erected, these losses could be reduced and avoided. The GDP decline is lower for countries that have a lower rate of cybercrime attacks. See Table 5.4.

In scenario 3, where we intend to estimate the losses due to strict liabilities and ex-ante regulations, there is a decline in GDP of almost all countries. This is due to the fact that imposition of such strict guidelines hinders countries from harnessing the AI benefits to the fullest extent. There may be a decline in investment of some regions, and so some countries may lag in adoption of AI technologies. Such frameworks also incur costs, reduce the pace of research and innovation, affect industrial prices, and may not be able to cope with the competition. The GDP losses from such regulations are high for Turkey (−2.66%), Australia (−2.58%), France (−2.50%), South Africa (−2.48%), and Argentina (−2.13%).

In countries such as Australia, Japan, Korea, India, Canada, Mexico, France, Russia, Turkey, and South Africa, the loss in scenario 3 is higher than scenario 2, meaning that the losses from ex-ante regulations and strict liabilities is higher than the loss from cybercrime attacks and data breaches. Thus it requires a harmonized framework that brings about a trade-off between ensuring data privacy and protecting AI systems from unwarranted cybercrime attacks to reduce the losses that incur from such problems, and ensuring that such regulations do not hamper investments, innovation, and research activities relating to development and adoption of AI systems.

5.4.2 Impact of policy shocks on output of countries/regions

The change in output of the three scenarios is shown in Figs. 5.1–5.3.

In scenario 1, it can be seen that most of the countries gain in manufacture of pharmaceuticals and medical products, light manufacturing, travel, extraction, telecommunication, machinery and equipment, etc. In Australia, the output of services sector increases, wheres that of the manufacturing sectors decrease. On the other hand, in Canada, the output of manufacturing sector increases, whereas that of the services sector decreases. In China all of the manufacturing sectors experience an increase in output, whereas it decreases in some of the service sectors. The decrease in some of the sectors in selected set of countries is due to the distribution of endowment commodities, such as land, labor, and capital from those sectors to that of the other sectors.

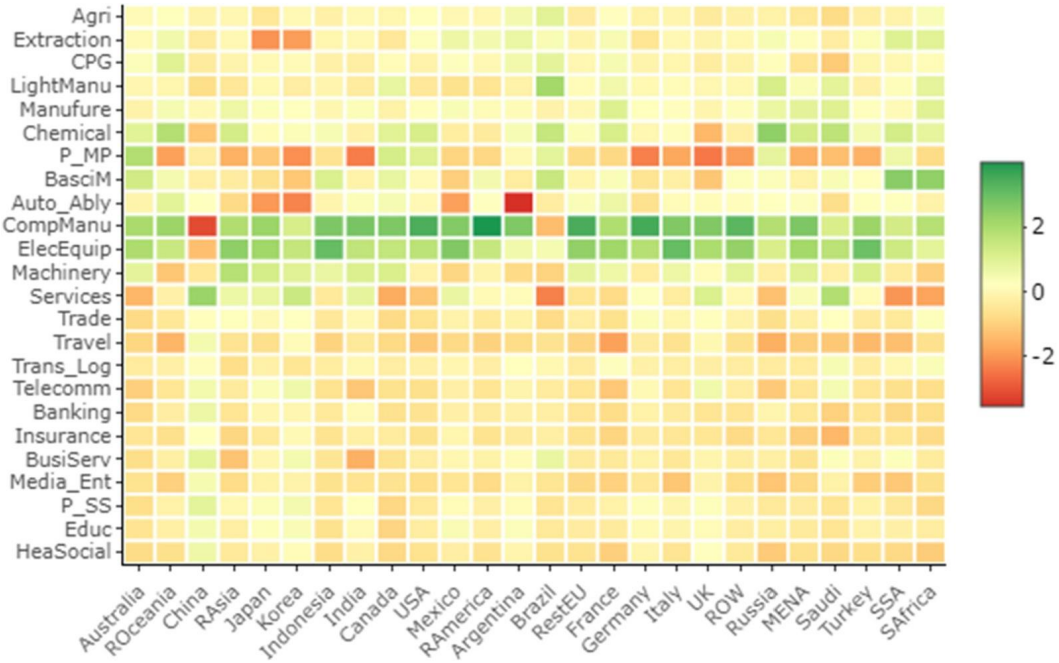


FIGURE 5.1 Impact of policy shocks on output of countries/regions; Scenario 1.

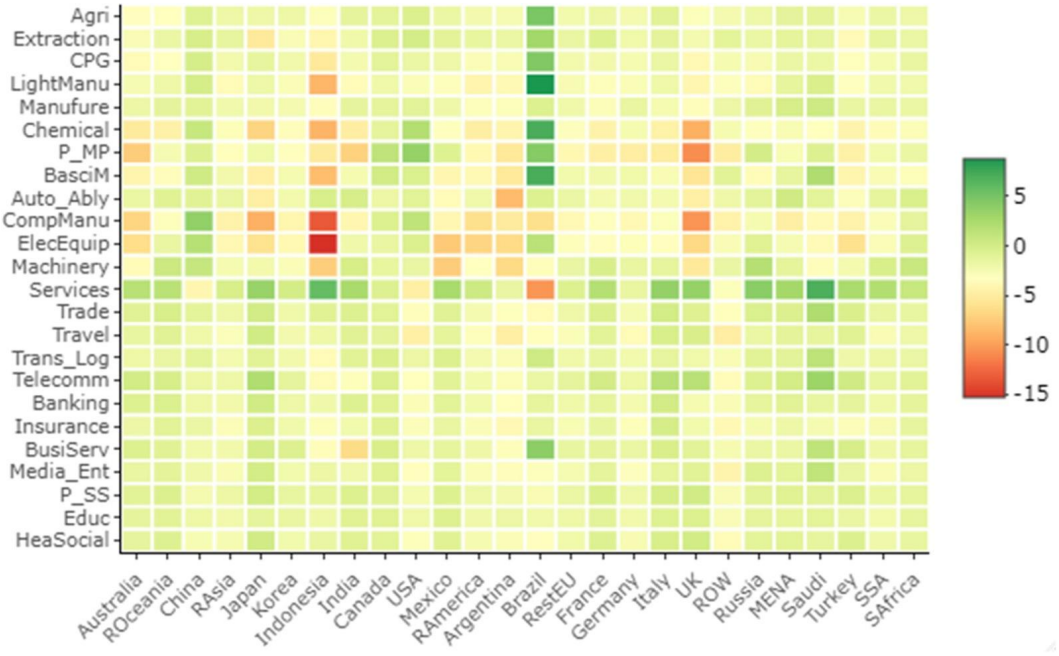


FIGURE 5.2 Impact of policy shocks on output of countries/regions; Scenario 2.

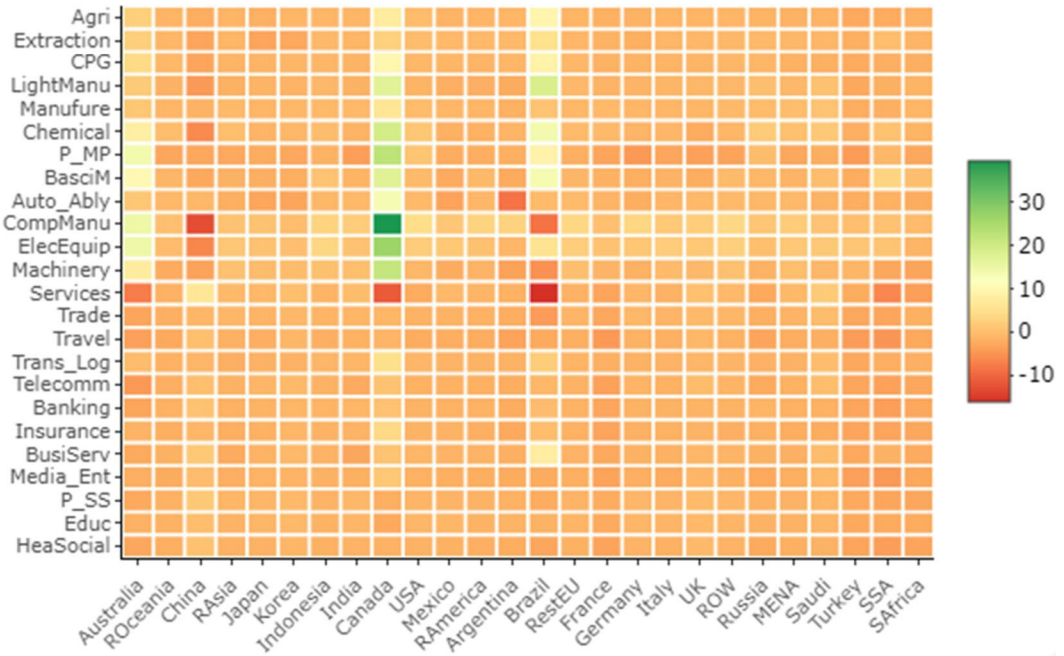


FIGURE 5.3 Impact of policy shocks on output of countries/regions; Scenario 3.

As in Fig. 5.2, most sectors experience a decline in output when shocked with cybercrime losses in scenario 2. Such losses could be avoided, and the gains from AI adoption could be maximized when proper regulatory regimes are erected.

Though the output of scenario 3 declines for most sectors in almost all the countries, the decline is lower than the decline in scenario 2. In some countries, such as China, Indonesia, Canada, and Brazil, the decline in output in scenario 3 is greater than scenario 2, meaning that the losses due to ex-ante regulations and strict liabilities is higher than the losses due to cybercrime attacks and data theft.

5.4.3 Impact of policy shocks on employment of countries/regions

When full adoption of AI is studied across all regions, it estimates an increase in the employment of unskilled labor across all regions. The highest increase is in Australia (10.1%), USA (7.25%), and the European Union

Table 5.5 Changes in employment of unskilled labor.

Countries	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)
Australia	10.10	−0.44	−4.24
Rest of Oceania	6.04	−0.58	−2.10
China	6.41	−1.90	0.87
Rest of Asia	5.01	−1.97	−1.16
Japan	6.44	0.22	−1.35
Korea	5.03	−1.72	−0.69
Indonesia	4.85	0.03	−1.10
India	2.91	0.11	−0.80
Canada	5.72	−0.96	−3.85
USA	7.25	−3.55	−1.75
Mexico	5.41	−0.92	−1.49
Rest of America	5.63	−2.06	−1.72
Argentina	5.88	−3.02	−2.22
Brazil	5.29	−3.85	−4.24
Rest of EU	6.73	−2.15	−1.54
France	3.93	−0.24	−2.70
Germany	5.94	−2.67	−1.31
Italy	5.23	−0.33	−1.35
UK	6.97	−0.02	−0.56
Rest of the world	5.74	−3.44	−1.09
Russia	2.61	0.74	−2.19
MENA	4.17	−0.13	−1.28
Saudi	1.86	1.12	−0.27
Turkey	5.54	−0.65	−3.13
SSA	5.85	−1.23	−2.94
South Africa	6.02	−1.36	−3.05

(6.73%). Though it is often argued that automation and technological enhancements from the adoption of Industry 4.0 and disruptive technologies, in our study the adoption of AI is estimated to increase the employment. This could be due to the fact that the automation increases the employment of skilled labor force. See Table 5.5.

In scenario 2, when cybercrime losses are considered, there is a considerable decline in employment of Brazil (−3.85%), the USA (−3.55%), Argentina (−3.02%), and Germany (−2.67%). These losses could be offset and

the gains in scenario 1 could be fully achieved by designing and adopting strict regulatory frameworks.

When losses from ex-ante regulations are fed as shock to the model as in scenario 3, the model estimates a considerable decline in the employment of Australia, Brazil, Canada, Turkey, and South Africa. In these countries, the decline in scenario 3 is greater than the losses in scenario 2. Thus these countries should identify and minimize interventions and strict liabilities that result from stringent regulatory measures and ensure that such regulations do not hinder the AI adoption.

5.4.4 Impact of policy shocks on export of countries/regions

In the first scenario, the export of almost all the countries increases, and it is predominant in India, Turkey, Argentina, and Japan. A sectoral analysis reveals that the increase is brought about by increase in exports of manufacturing industries. On the other hand, there is a decrease in export of Australia, because there is an increase in domestic consumption, which could be due to increase in wealth of the consumers.

In the second scenario, the export of most of the countries decline; Indonesia, Japan, India, UK, Turkey experiences a significant decline in aggregate exports. The decline is due to the loss in the output brought about by cybercrime attacks. Strict liabilities and regulatory frameworks could reverse this loss. Countries that have lower rate of cybercrime attacks experience a gain in exports. Those countries that have higher rate of cybercrime attacks produce a lower economic output and so may depend on imports from those countries that have a lower rate of data theft.

In scenario 3, the export of most of the countries decline: China, India, Japan, and Korea. Those countries that have higher estimates of losses incurring from strict liability framework as a percentage of GDP, experience a decrease in exports and other countries with lower loss estimate experience an increase in exports. See Table 5.6.

5.4.5 Impact of policy shocks on import of countries/regions

In scenario 1, where it is assumed that all countries adopt AI in all sectors, there is an increase in imports among most of the countries. Countries such

Table 5.6 Change in exports.

Countries	Base Value (in million USD)	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)
Australia	272,131	−10.80	−5.75	10.60
Rest of Oceania	64,910	8.44	−4.72	−0.88
China	2,408,744	9.03	5.69	−16.80
Rest of Asia	2,029,627	8.22	−3.31	−1.41
Japan	843,535	13.50	−12.00	−2.57
Korea	713,906	9.83	−3.72	−2.18
Indonesia	220,358	7.21	−14.00	−0.22
India	459,663	20.20	−10.20	−3.51
Canada	493,295	4.99	−0.74	17.40
USA	1,892,909	−4.53	8.27	4.28
Mexico	452,007	10.00	−4.98	−1.48
Rest of America	467,173	7.05	−3.63	−0.73
Argentina	75,018	15.20	−4.99	−0.43
Brazil	269,937	5.16	25.20	43.90
Rest of EU	6,080,126	4.66	−2.81	−0.23
France	20,001	7.21	−3.06	−0.91
Germany	97,858	7.32	−2.49	−1.26
Italy	131,845	5.47	−3.24	−0.51
UK	658,723	2.83	−10.10	−2.67
Rest of the world	862,421	5.10	−1.83	−0.92
Russia	546,065	6.59	−5.08	0.45
MENA	1,069,971	5.58	−2.84	−1.08
Saudi	360,495	4.97	−2.76	−1.14
Turkey	208,325	16.90	−7.82	−2.18
SSA	357,509	2.61	−2.81	1.46
South Africa	116,985	6.08	−3.30	0.49

as Australia, USA, Sub-Saharan Africa, and China experience a significant increase in imports. See Table 5.7. This is due to an increase in import of manufacturing sectors in some countries, where an increased output may demand more capital goods and raw materials. Whereas in other countries, there is an increase due to increase in consumption.

In scenario 2, there is a significant decline in the import of Brazil, USA, and China, as these countries experience higher losses in GDP due to cyber-

Table 5.7 Change in imports.

Countries	Base Value (in millions USD)	Scenario 1 (%)	Scenario 2 (%)	Scenario 3 (%)
Australia	267,659	19.20	2.50	−10.00
Rest of Oceania	77,630	1.85	1.50	−1.46
China	2,093,751	7.47	−4.13	5.93
Rest of Asia	1,944,355	4.08	−1.42	−0.84
Japan	889,156	2.96	3.47	−0.82
Korea	617,480	4.53	−1.39	−0.56
Indonesia	194,033	4.16	4.85	−1.63
India	548,038	−0.30	2.65	−0.62
Canada	520,500	5.97	−0.90	−4.76
USA	2,791,386	11.40	−6.98	−3.70
Mexico	417,780	4.06	−0.50	−0.93
Rest of America	485,681	2.97	0.25	−1.11
Argentina	86,331	−0.17	1.09	0.43
Brazil	261,884	5.70	−13.60	−19.70
Rest of EU	5,920,725	5.10	−1.40	−1.16
France	21,291	3.31	−0.08	−2.21
Germany	94,371	4.41	−2.10	−0.85
Italy	117,803	4.37	−0.73	−0.68
UK	825,183	5.97	2.88	0.24
Rest of the world	770,065	4.52	−2.95	−0.65
Russia	323,879	1.71	4.50	−3.78
MENA	1,019,416	3.06	1.15	−0.93
Saudi	171,911	2.47	2.61	−0.12
Turkey	252,628	1.56	1.30	−2.13
SSA	355,966	7.58	0.74	−5.61
South Africa	104,636	4.28	0.34	−3.24

crime attacks. Some countries, such as Indonesia, that experience a higher rate of cybercrime are estimated to experience an increase in imports. This is due to the increase in imports of such countries to meet the domestic consumption demands on account of decline in industrial output. Countries, such as the UK, Japan, and Australia, that experience a lower rate of cybercrimes, are expected to increase their imports due to increase in economic activity.

In scenario 3, all the countries experience a decline in import. The decline is significant in Brazil, Australia, Canada, South Africa, and Russia, as these countries experience a higher loss (as a percentage of GDP) due to imposition of strict liabilities and restrictive legal frameworks.

5.5 Conclusion

We attempted to discuss the potential gains from full adoption of AI technology across many countries /regions of the world; the losses that may occur due to the lack of a strong ethics framework to protect AI from cybercrimes and the potential challenges in strict regulations that may be aimed at enforcing ethics in AI, but end up harming the economy as a collateral damage. The global economic model we use (GTAP) is capable of capturing the inter-sectoral linkages between different regions.

The impact of AI across the regions at the sector-level are analyzed in three scenarios: the first considers complete adoption of AI across all the countries to quantify the economic benefits of AI athwart all industry sectors and its positive impact on variables such as output (GDP), productivity, trade, investments, innovation, consumer welfare, and increase in worker wages; the second considers the gains that each country can make from adopting ethical frameworks and standards to curb illegitimate usage of such technologies and to regulate their adoption; and the third scenario estimates the losses that each country/region would endure due to the imposition of strict liabilities and restrictive legal norms, which in turn could curb countries from leveraging the true potential of AI.

We find that there will be a gain in GDP of all the regions when full adoption is assumed in scenario 1. Australia, China, and USA would gain the most to about 6.77%, 6.38%, and 6.1% of GDP, respectively. In scenario 2, the study estimates a loss in GDP to all countries when cybercrimes and data thefts are left unprotected. Argentina, Brazil, USA, and Germany experience higher losses as cybercrime rates as a percentage of their GDP is higher in these countries. Such losses could be mitigated by imposing and strategizing legal frameworks to govern the development and usage of AI systems. In scenario 3, when the model is shocked to estimate the losses due to imposition of strict liabilities and ex-ante regulations, Argentina, France, Australia,

Turkey, and South Africa experience higher losses. In some countries, the loss in GDP in scenario 3 is higher than in scenario 2, meaning the losses that incur from restrictive legal practices is higher than the losses from cybercrime attacks. The rate of employment increases in scenario 1 against the typical argument and concern raised by most experts stating that automation could lower employment rates, thanks to much greater economic expansion than the extent to which capital may substitute labor due to the lower labor intensity of AI.

Overall, the economic benefits from AI outweigh the costs and losses that could incur from cybercrimes, data thefts, strict liabilities, and employment. Strict regulations that are originally imposed with a purpose of preventing data privacy have adverse impacts on innovation and implementation of AI systems in some countries. Thus a harmonized framework is the need of the hour.

Acknowledgment

The authors thank Dr. Hosuk Lee-Makiyama, Sudha Varadhan, Ashwini Ramu, Divyayudha Khire, Adems George, and Sampriti Sharma for their comments, inputs, and assistance.

References

Aguiar, A., Chepeliev, M., Corong, E., McDougall, R., van der Mensbrugghe, D., 2019. The GTAP data base: version 10. Journal of Global Economic Analysis 4 (1), 1–27.

Arcesati, Rebecca, 2021. Lofty principles, conflicting incentives: AI ethics and governance in China.

Bendett, S., 2020. Putin Urges Ai Limits — But for Thee, Not Me? Defense One. https://www.defenseone.com/ideas/2020/12/putin-urges-ai-limits-thee-not-me/170458/.

Bughin, Jacques, Seong, Jeongmin, Manyika, James, Chui, Michael, Joshi, Raoul, 2018. McKinsey Global Institute.

Cain, Nathan, Wahid, Usman, Allen, Leanne, Ost, Isabell, Peart, Alex, 2020. Spotlight on AI regulation.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., Malhotra, S., 2018. Notes from the AI Frontier: Insights from Hundreds of Use Cases. A McKinsey Global Institute Study.

Commonwealth Scientific and Industrial Research Organisation, 2019. Artificial Intelligence Australia's Ethics Framework.

CSIS, 2014. Net Losses – Estimating the Global Cost of Cybercrime, Economic Impact of CyberCrime.
Davidson, Leah, 2019. Narrow vs. General AI: What's next for Artificial Intelligence?.
Dean, Digrande, Field, Lundmark, O'day, Pineda, Zwillenberg, 2012. The Connected World, the Internet Economy in the G-20. BGC Report.
ELSI, 2021. The Japanese Society for Artificial Intelligence Ethical Guidelines.
European Commission, 2020. White paper on Artificial Intelligence - a European approach to excellence and trust.
Evas, 2020. Civil liability regime for artificial intelligence. European added value assessment. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654178/EPRS_STU(2020)654178_EN.pdf.
Gal, D., 2020. China's Approach to AI Ethics. Nesta.
Global Legal Insights, 2021. AI, Machine Learning & Big Data Laws and Regulations 2021 India.
Hertel, T.W., 1997. Global Trade Analysis: Modeling and Applications. Cambridge University Press, Cambridge, United Kingdom.
Hertel, Thomas, Hummels, David, Ivanic, Maros, Keeney, Roman, 2004. How confident can we be of CGE-based assessments of free trade agreements? Economic Modelling 24, 611–635. https://doi.org/10.1016/j.econmod.2006.12.002.
Hunt, Mia, 2020. Indonesia publishes AI Strategy. In: Global Government Forum.
Hwang, Park, 2019. The threat of AI and our response: the AI charter of ethics in South Korea. Asian Journal of Innovation and Policy 9 (1), 056.
Lee-Makiyama, Hosuk, Narayanan, Badri, 2020. Economic Costs of Ex-ante Regulations. ECIPE Journal. https://ecipe.org/publications/ex-ante/.
McKinsey Global Institute, 2019. Artificial Intelligence in the United Kingdom: Prospects and Challenges.
Menon, Rekha, Roy, Pradeep, 2021. Accenture Report. Rewire for growth.
Mondaq, 2020. India: Self-Regulation in Artificial Intelligence: an Indian Perspective.
Morgan, Steve, 2020. Cybercrime to Cost the World $10.5 Trillion Annually By 2025. Cybercrime Magazine. https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/.
Observation Research Foundation, 2019. Artificial Intelligence in Africa's healthcare: Ethical considerations.
OECD, 2021. 6. National policies for Artificial Intelligence: What about diffusion? The Digital Transformation of SMEs.
PDPC, 2019. Model Artificial Intelligence Governance Framework.
PWC, 2018. The macroeconomic impacts of artificial intelligence.
Swiatkowsa, Joanna, 2020. Tackling Cybercrime to Unleash Developing Countries' Digital Potential. Background Paper 23.
U.S. Department of Defense, 2020. DOD Adopts Ethical Principles for Artificial Intelligence.

This page intentionally left blank

PART 2

AI assurance methods

This page intentionally left blank

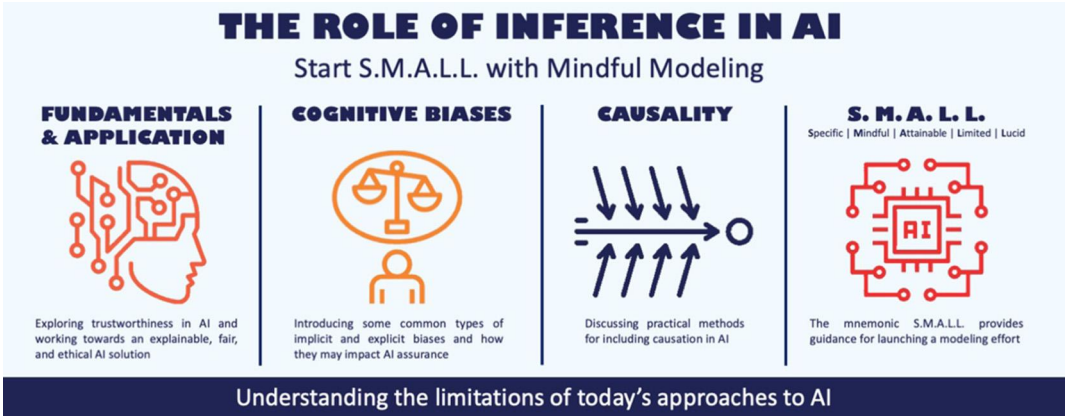
6

The role of inference in AI: Start S.M.A.L.L. with mindful modeling☆

Jay Gendron[a,d] and Ralitsa Maduro[b,c]

[a]*Model the Cause, Chesapeake, VA, United States* [b]*Sentara Healthcare, Virginia Beach, VA, United States* [c]*Virginia Wesleyan University, Virginia Beach, VA, United States*

Graphical abstract



Abstract

This chapter explores trustworthiness in AI and penetrates the black-box opacity through explainable, fair, and ethical AI solutions. AI remains a spirited topic within academic, government, and industrial literature. Much has occurred since the last AI winter in the early 1990's; yet, numerous sources indicate the

☆ S.M.A.L.L. is an acronym for Specific-Mindful-Attainable-Limited-Lucid.

[d] J. Gendron published under the affiliation Model the Cause to comply with releasability policies in his role as a data scientist. Model the Cause is an organization that promotes Mindful Modeling approaches through volunteerism and training.

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00019-6

initial successes solving problems like computer vision, speech recognition, and natural sciences may wane — plunging AI into another winter.
Many factors contributed to advances in AI: more data science courses in universities producing data-science capable graduates, high venture capital funding levels encouraging startups, and a decade of broadening awareness among corporate executives about AI promises, real or perceived. Nonetheless, could sources like Gartner be right? Are we approaching another AI winter? As the world learned during the COVID-19 pandemic, when we find ourselves in a crisis, focusing on the fundamentals can have a powerful effect to easing the troubles.
As AI makes history, it relies on progress from other domains such as data availability, computing power, and algorithmic advances. Balance among elements maintains a healthy system. AI is no different. Too much or too little of any elemental capability can slow down overall progress. This chapter integrates fundamental ideas from psychology (heuristics and bias), mindfulness in modeling (conceptual models in group settings), and inference (both classical and contemporary).
Practitioners may find the techniques proposed in this chapter useful next steps in AI evolution aimed at understanding human behavior. The techniques we discuss can protect against negative impacts resulting from a future AI winter through proper preparation: appreciating the fundamentals, understanding AI assumptions and limitations, and approaching AI assurance in a mindful manner as it evolves. This chapter will address the fundamentals in a unifying example focused on healthcare, with opportunities for trustworthy AI that is impartial, fair, and unbiased.

Keywords

AI assurance, cognitive bias, intrinsic bias, machine learning, Bayesian statistics, causal inference, systems validation

Highlights

- Continued progress in artificial intelligence (historically deterministic and structured problems) will broach the boundaries to understanding human behavior
- Leading researchers in artificial intelligence highlight the need (and potential) from incorporating causal inference into deep learning
- Capturing causality in systems requires not only data but also conceptual understanding using established practices from systems dynamics and the social sciences

- The Start S.M.A.L.L. approach is a practical method to aid practitioners with mindful modeling in various domains and help discover bias sources

6.1 Real wisdom on artificial intelligence

Artificial Intelligence (AI) attained a new high-water mark, expanding the domain not just in mathematical optimization techniques but also statistical approaches and exploiting access to large amounts of data. AI is a big deal that will continue to shape the next half century. Media coverage highlights promises and cautions, but what are lay people and analytic practitioners to make of the hype and the concerns?

Much occurred after the last AI winter (a period of reduced resources and progress) and AI boomed in popularity since the mid-2000s, first within academia and now within government and corporate enterprises.

> *The new spring in AI is the most significant development in computing in my lifetime. Every month, there are stunning new applications and transformative new techniques. But such powerful tools also bring with them new questions and responsibilities (Vincent, 2018).*
> — Sergey Brin

Many factors contributed to this boom: cost effective distributed computing power, numerous personalized data sources, data science courses in universities producing more data science capable graduates, ample access to venture capital encouraging startup businesses, and a decade of broadening awareness among corporate executives about AI promises to boost profits and improve operations. Nonetheless, could experts at Gartner (Kinsella, 2017) be right? Are we approaching the next AI winter?

Like any technology, AI will experience its natural plateaus on the technological S-curve (Sterman, 2000). The last AI winter was one such plateau. Right now, AI is somewhere in the growth phase. Has the boom consumed the "low-hanging fruit" of this era? By low-hanging fruit the authors mean:

- automated work agents like chatbots, data pre-processing, and automated machine learning

- improved predictive power in physical systems or problems having an underlying structure, like computer vision, speech recognition, and the natural sciences
- successful penetration into commercial markets with deep learning solutions

As the world learned from the COVID-19 pandemic, during a crisis, focusing on the fundamentals can have a powerful effect in promoting progress. AI relies on many elemental capabilities: data availability, computing power, algorithmic advances, and fundamentals. Like many comparable systems in the world, equilibrium is required to maintain a healthy system. This is no different in AI. Too much or too little of any elemental capability slows down overall progress. This book explores trustworthiness in AI and penetrates the black-box opacity through explainable, fair, and ethical AI solutions. This chapter highlights fundamentals found throughout science that continue to garner attention in the literature and could support continued progress in AI evolution:

- cognitive bias and heuristics bias from a psychological perspective, based on the work of Kahneman and Tversky
- mindful modeling approaches to complement increased capabilities for data mining, based on the work of Sargent (conceptual modeling) and Luna-Reyes (group model building)
- the roles of inference and causality, based on the work of Sterman (system dynamics) and Pearl (causal inference)

As the low-hanging fruit disappears, this chapter identifies opportunities to undertake and solve the next evolutionary cycle of problem sets. Impacts from the next AI winter may be blunted through proper preparation: appreciation for the fundamentals, consideration of the importance of AI assurance and how it relates to AI bias, and discussion of inferential methods to support AI assurance. This chapter will address the fundamentals in a unifying example focused on healthcare, with opportunities for trustworthy AI that is impartial and fair (free of conscious and unconscious bias).

The chapter is organized into two parts: fundamentals and application. **Section 6.2** provides important fundamentals for AI practitioners: decision-

making, heuristics, and bias. It provides a relevant summary of psychological constructs to support later chapter sections. **Section 6.3** lays out two other topics fundamental to AI: the discipline of modeling and a summary of inferential methods. AI assurance design in **Section 6.4** provides a working scenario focused on how medical errors are causing a healthcare crisis. The section integrates the fundamentals along with the emerging literature on AI assurance using the S.M.A.L.L. framework for mindful modeling. **Section 6.5** provides a short summary and takeaways, and **Section 6.6** provides further reading suggestions for those interested in learning more.

6.2 Fundamentals: decision-making, heuristics and cognitive biases

> *In the context of a society whose dominant elements justify their position by arguing the genetic inferiority of those whom they dominate, it is hard to be neutral (David, 2001).*
> — Richard David, MD, Stroger Hospital of Cook County, Chicago

There are two categories of decision-making theories: philosophically normative theories (e.g., how people should make decisions) and the empirically supported descriptive theories (e.g., how people do make decisions; (Beresford and Sloper, 2008)). This section reviews only the descriptive decision-making theories, as they apply directly to medical decision making as a use case where AI has penetrated the market. Most such theories subscribe to the two-system, dual-process view of decision-making (Hogarth, 2002; Kahneman, 2011; Sloman, 1996).

The leading dual-process theory of decision-making is prospect theory, developed by Daniel Kahneman and Amos Tversky (1979). **Prospect theory** was developed as an alternative to the previously dominant Expected Utility Theory (Neumann and Morgenstern, 1953) that asserted human rationality is equivalent to a specific mathematical model, interpreting every choice as the maximization of an individually tailored, real-valued utility function with specific mathematical properties (e.g., monotonicity). Prospect theory is an alternate theory of decision making under uncertainty and risk that better aligns with both experiment and experience. It explains how individuals without extensive education or experience at-

tempt to make optimal choices without assuming they are naïve utility maximizers. A main premise of the theory is that people use **heuristics** (quickly-applied rules of thumb that guide behaviors without the need for deep processing) in order to make decisions under uncertainty. Heuristics are subjective and individualized, such that a person's use of heuristics is based solely on what they know at the point of decision-making. Heuristics do not suppose additional information-seeking before a decision is made (Kahneman, 2011). Similar to habits, heuristics are developed based on individual experiences.

6.2.1 Dual-process model of decision-making

The dual-process model of decision-making describes two different modes or processing styles under which decisions are made. These are referred to as System 1 and System 2. **System 1** thinking is intuitive, operates automatically and quickly, requires little effort, and is often strongly influenced by emotions. An example of System 1 thinking is stereotyping (Kahneman, 2011); humans often make impressions of others only seconds after meeting them. Although hostile stereotyping can be harmful, the overall process of categorizing people (e.g., angry and dangerous versus friendly and harmless) is adaptive, useful, efficient, and fast. The need to categorize and make judgments of situations quickly, without using more complex and effortful reasoning, is the basic function of System 1 (Evans, 2008).

Unlike System 1, **System 2** thinking is deliberate, analytical, and requires greater cognitive effort and attention (Kahneman, 2011); performing a mathematical calculation is an example of System 2 thinking. There is little use of intuitive thought when performing complex calculations. Instead, a person focuses their attention on the problem and engages in the cognitive efforts of following general rules. Cognitive strain can have a negative impact on the performance of System 2. Poorly designed processes, physical spaces, or software can further heighten cognitive strain throughout the day.

Both System 1 and System 2 are susceptible to biases and errors. In certain situations System 1 may perform well in collecting relevant information; however, due to our bias towards ignoring evidence we dislike (see

confirmation bias in Section 6.2.3 below), System 2 may make a mistake: I went for a run this morning, so it is fine to have a big slice of cake now. Conversely, System 1 (which commonly uses heuristics) may have gathered biased evidence so even if the System 2 processes run accurately, the outcome may be incorrect. Lastly, neither System 1 nor System 2 are superior to one another. Kahneman (2011) clearly states that System 1 is not often error-prone and System 2 is not always correct. Research comparing the accuracy of prediction using System 1 heuristics (take-the-best, tallying, and minimalist) to two predictive analytic strategies (linear regression and naïve Bayes) found that System 1 strategies outperformed the complex System 2 strategies when there was limited initial data (Hertwig and Pachur, 2015).

6.2.2 Error and bias in medical decision-making

It is important to note that System 1 and System 2 thinking do not operate in isolation from one another or other influences. Decision-making theorists agree that the two systems lay at the ends of a continuum and that people are often making decisions by employing both analytical and intuitive thought. The location within the continuum on which a healthcare decision-maker falls would vary based on their prior life experiences or learning; cognitive and emotion regulation abilities; and implicit and explicit biases. System 1 thinking can provide great benefits in terms of medical decision making. It allows providers to perform in situations potentially harmful to their patients. For example, when a patient exhibits life-threatening physiological symptoms (e.g., rapid change in vital signs), the provider's reaction aimed at saving the patient's life is quick and intuitive, and there may be strong emotions related to the process (e.g., anxiety). Furthermore, the effortless nature of System 1 thinking empowers providers with the cognitive ease that makes everyday tasks less exhausting. In other words, System 1 thinking allows providers to use their cognitive resources for more critical dilemmas regarding patient needs. Unfortunately, System 1 driven provider thinking may not be optimal when the provider's reaction to ambiguous symptoms is too emotional, non-rational, and in itself harmful to the patient (e.g., unnecessary treatment with antibiotics for a viral infection, due to provider anxiety brought on by demanding patients).

Also, potential issues may arise when a provider's System 1 response is maladaptive, arising from a problematic learning history involving explicit and implicit bias, for example.

Similarly, a healthcare provider engages in System 2 thinking if they are actively considering the best treatment option or diagnosis for their patient. This process would require cognitive effort to make a decision based on numerous clinical parameters. Sometimes providers may not have the cognitive capacity to engage in System 2 thinking because of uncontrollable external conditions. For example, System 2 thinking may be too effortful for providers already exhausted or under a heavy cognitive load for other reasons. In addition, providers may not prefer System 2 for decisions that are already well practiced (i.e., habits acquired and cemented in place by experience). Research shows that providers can improve effortful decision-making, such as accurate diagnosis of heart disease, by employing "fast and frugal" decision trees which follow an evidence-based algorithm (Green and Mehr, 1997). AI solutions on the marketplace now offer such algorithms and more.

The high prevalence and negative consequences of biases (explicit and implicit) and errors in decision-making among US healthcare providers are well-documented (Aronson et al., 2020; Johnson et al., 2004; Phelan et al., 2014). Examples range from unintentional cognitive errors such as administering the wrong medication (Anderson, 2019) to racial bias from differences in emergency department pain management (Todd et al., 1996). Medication errors, the most common medical errors, are the leading cause of hospital morbidity and mortality (in the US and internationally) and continue to increase in frequency (De Vries et al., 2008; Jha et al., 2013). Indeed, the National Pharmacy Association quarterly medication safety update report stated medication errors account for 66 percent of all errors reported (National Pharmacy Association, 2021). As called for by Panagioti et al. (2019), efforts should focus on understanding preventable patient harm, a continual and serious problem across medical care settings. According to their meta-analysis, mitigation of major sources of preventable patient harm are priority areas of future work. Healthcare AI solutions claim to provide such mitigating solutions.

AI-enabled healthcare decision-making tools represent a growing market. Hundreds of vendors now offer predictive and prescriptive analytic tools under the terms machine learning, AI, or cognitive machine. The promise of all these "smart" solutions is to help providers within a healthcare organization make smarter and faster decisions: promising the speed of System 1 and the reliability of System 2 thinking. Unfortunately, medical errors and hospital mortality rates continue to remain prevalent, despite the fast adoption of AI in healthcare. In retrospect, many AI practitioners now realize the historical databases used to develop these algorithms tend to automate and propagate the biases of the decision-makers who created them. Of course, algorithms are not the only source of bias (e.g., trained on biased data sets). Biases propagate by the people who develop AI algorithms (e.g., implicit racism, sexism, ageism, or ableism) and indirectly to products. Lastly, users present bias as well. Part of the issue why AI solutions may not be as efficacious as expected is that the providers who use them fall prey to automation complacency whereby they accept the recommendations of the AI tool and stop investigating the conditions any further (especially when the AI algorithm predicts everything is "normal"; (Parasuraman and Manzey, 2010)). Therefore, much of AI assurance focuses on uncovering algorithms that reinforce and perpetuate human biases.

6.2.3 Implicit and/or explicit: bias in AI practitioners and AI models

With the benefits of AI also come the threats of implicit and explicit bias whenever AI is used. Implicit bias is a psychological process in which a person's unconscious beliefs and attitudes affect their behaviors, perceptions, and judgments in ways unaware to themselves, and typically, unable to control. Explicit bias refers to the attitudes and beliefs we have about a person or group on a conscious level. Next, we introduce some common types of implicit and explicit biases and how they may impact AI assurance.

Ableism is defined as a network of beliefs, processes, and practices that produce notions of a perfect human body within the human species (Campbell, 2001). Algorithms can produce a bias towards enabling more perfect human cognition and decision-making which in turn stigmatizes those

with different abilities. Shew (2020) coined the term ***technoableism***, which is a particular strain of ableism in the context of imagination, technology, and bodies. As Shew points out, AI solutions such as autonomous vehicles, companion robots, and caretaker robots are marketed as ways to remedy problems among individuals who are aging or have a disability. The author argues that focusing on the individual's problem ignores society's responsibility for creating better planning and infrastructure within communities where older or disabled individuals can leave their homes without need for AI-driven, in-home robots to help with loneliness and isolation. Shew (2020) further discusses the problem with AI designers who create solutions but are not themselves part of the disability community. As Shew states, "These designers who are usually ignorant of the larger history of disability, often reinforce ableism in their design, further stigmatizing and marginalizing disabled people through monitoring or tracking or decision-making by proxy" (2020). The AI technology that developers create in an effort to achieve top performance is rooted in ableism via predictive models based on having groups designated as superior and inferior (Council of Europe Directorate General Human Rights and Rule of Law, 2019; Krupiy, 2020).

Affinity bias is the unconscious tendency to gravitate toward people who look like us, have the same beliefs, and come from the same background. Moreover, due to affinity bias, we may avoid or even dislike people who are different from us. In AI, machine learning models developed by white males that consume data not representative of the population can show an affinity towards the white and male. At least one study provided evidence that affinity bias is present in human-robot interactions. Specifically, avatar appearance impacted the preferences of humans asked to sort resumes of gender and skin-tone varying avatars (Trainer et al., 2020). In addition, research shows there is an own-race bias in face recognition models where deep learning networks demonstrated a strong tendency to focus on selected facial regions for a particular race (Nagpal et al., 2019). Indeed, a wealth of recent literature outlines examples of racism and sexism in AI. Some noteworthy examples are Tay, the anti-Semitic Twitter chatbot launched by Microsoft (Beran, 2018), incorrect and racially-biased recidivism prediction AI (Angwin et al., 2016), and the commercial face recognition software which

was efficient in identifying lighter-skinned males but not when detecting darker-skinned females (Lohr, 2018). Although with the chatbot example, the racism was likely implicit on the side of the developers, some raise concerns that developers make explicitly race- and gender-biased decisions to maximize profit (Buolamwini and Gebru, 2018; Hong and Williams, 2019; Noble, 2018).

Ageism bias is prejudice or discrimination on the grounds of a person's age. Own-age bias was present in AI algorithms according to the analysis of Nagpal and colleagues (2019) across multiple deep learning networks. This bias is documented in other evaluations of big data approaches as well. In a recent review of academic literature, instances of both implicit and explicit ageism existed in algorithms, intelligent systems, and big data (Rosales and Fernández-Ardèvol, 2019). Specifically, the authors note that when it comes to age prediction and smartphone data analysis, the older population is often not included in big data approaches. In the words of the authors "older people are invisibilized by, for instance, not controlling the capacity of the sample to represent the studied population." Age-biased samples (samples based on social media and smartphone users) produce age-biased tools. Often, algorithms only manage to distinguish between younger cohorts, as age predictions tend to work better for younger cohorts (Culotta et al., 2016; Liao et al., 2014; Nguyen et al., 2013; Peersman et al., 2011).

Attribution bias is a type of unconscious cognitive bias that refers to the systematic errors of making more favorable assessments of the behaviors of those in your "in group" while applying stereotypes and judging people more harshly when they are in your "out group" (Nalty, 2016). Novel research suggests that attribution bias can be present when humans anthropomorphize robots and assign them a social categorization (Haring et al., 2018; Kuchenbrandt et al., 2013). In particular, form function attribution bias research shows that just as with their person-to-person interactions, humans use cognitive shortcuts based on their visual perception of a robot, instead of objectively evaluating its functionalities during an interaction (Haring et al., 2018).

Confirmation bias is one of the most common threats to AI assurance. It is a cognitive bias commonly found in humans and is the tendency to attend

to similarities (Gilovich et al., 2002). Confirmation bias is often discussed as an overarching category of other similar biases in AI and can be described as a phenomena where one seeks evidence that is consistent with their current hypotheses and interprets evidence only when supporting their point of view. A common example of this bias in AI is the increase in ads for certain products that a customer has already purchased (Chou et al., 2017). If confirmation bias is ignored, it can lead to insights that are not well grounded in users' experiences and therefore it can stifle innovation (Butler and Roberto, 2018). Useful strategies proposed by researchers to combat confirmation bias include, but are not limited to, considering how the opposite of your belief might be true and forcing yourself to record discordant data (Darwin's two notebooks strategy) (Garvin, 2003; Lord et al., 1984).

Conformity Bias, or the human's tendency to act in a way that allows them to fit in, is also found in AI research. A survey of IT professionals who self-reported the likelihood that cognitive biases impacted their development of AI systems, showed that conformity bias was the most influential, according to respondents (Cazes et al., 2021). Interestingly, not only developers of AI products but also AI algorithm users are impacted by conformity bias. In a study examining provider decision making in AI-supported second opinion, settings showed that physicians preferred to conform to the decision of previous doctors in comparison to the AI algorithm decisions (Cabitza, 2019). An interesting and innovative type of conformity bias is a construct by Cheshire (2017) called "loopthink". The author posits that the old conformity bias driven "groupthink" (i.e., a phenomenon that is characterized by the act of making decisions as a group in a way that conforms to norms and discourages individual responsibility) will translate into loopthink when AI algorithms make decisions. He outlines two types of loopthink: **weak loopthink**, the "intrinsic inability of a sophisticated computer to redirect executive data flow as a result of its fixed internal hardwiring, uneditable sectors of its operating system, or unalterable lines of its programming code." In other words the computer is responding in a way that resembles a stubborn individual who is refusing to listen; and also **strong loopthink**, "an artificial intelligence's suppression, as a result of internalization of the ethical framework of its collective, of internal data processing pathways that,

if considered, could redirect executive output". An example of strong loop-think is a self-driving car computer which may swerve towards a child on the road instead of towards a billboard of three children as it is programmed to minimize the death toll in the event of unavoidable harm. This scenario errs from the images of three children on a billboard compared to one actual child on the road.

6.3 Fundamentals: yearning to make sense of the world through models and inference

> *Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity (Box, 1976).*
> — George Box

Well before machine learning, or theoretical statistics, or even before Newton and Leibniz developed their calculus, **modeling** existed. Modeling may be as old as thought itself (Frigg and Hartmann, 2020; Hodges, 2020). The introductory paragraphs of this section provide a glimpse into the relationship among contemporary modeling techniques and serve as a scene setter for practitioners as well as readers interested in AI.

Modeling is an immense domain encompassing algorithms and numerical methods. Modeling is also a way of thinking, a philosophy, by approaching a complex reality and simplifying it through assumptions and limitations to communicate a larger idea. The Princeton psychologist, Tania Lombrozo, was interviewed (March 2021) on why models matter and how our brain makes sense of questions. In discussing what drives the effects we see in a complex world, Lombrozo summarized models in a useful way: they relate our observations (high variability) to a more generalized underlying structure and predict how things may go in the future (Vedantam, 2021).

AI is a particular discipline that relies on modeling and makes use of techniques designed to mimic human responses to situations and deci-

Table 6.1 Machine Learning — Three general categories of machine learning.

Machine Learning Category	Description
Supervised Learning	algorithms learn underlying correlations and rules from data observations labeled with ground truth to predict future, unseen events
Unsupervised Learning	algorithms find underlying patterns in the data to provide insights for other modeling, such as supervised learning
Reinforcement Learning	algorithms learn behaviors based on reward and are used in learning games like chess and developing autonomous vehicles

sions. They may be as simple as a rules-based system or as involved as an ensemble of complex algorithms. A hierarchy exists within the field: AI is a superset that contains the discipline of **machine learning** while **deep learning** is a subset specialization within machine learning (Mokli et al., 2019).

Machine learning captured the hearts and minds of the public after data science became "the sexiest career of the 21st century" (Davenport and Patil, 2012). Machine learning relies on statistical and mathematical techniques to "learn" a phenomenon based on data. There are hundreds of algorithms used in machine learning, but the breadth of the work is often captured by three categories (Alpaydin, 2010; Hastie et al., 2001; Murphy, 2012). These are summarized in Table 6.1.

Mokli et al. provide a summary of the more common machine learning algorithms and models labeled according to their category of learning method. The first type of machine learning listed is linear regression. A representative example of this common type of (supervised) machine learning is illustrated in Fig. 6.1.

This example shows the simplest type of linear regression in two dimensions. The dots represent the actual data sample. The solid line running through the sample is the algorithmic-based **prediction** of the underlying system modeled by regression. The model (solid line) enables prediction of future, unseen observations within the range of the x-axis (in this case, *TV*). Linear regression is not limited to two dimensions. When working in multiple dimensions, hyper planes replace lines and exist within a hyperspace.

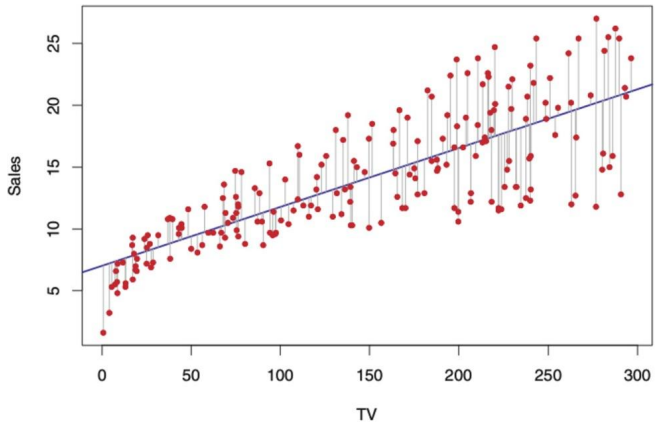


FIGURE 6.1 Example of Linear Regression — Among the more common machine learning techniques is linear regression to learn the linear trend, shown by the solid line, among the data, shown as dots (James et al., 2013).

Deep learning is a subset of machine learning and is a booming growth space in academia and industry (Benjamens et al., 2020; Lundervold and Lundervold, 2019). It boasts a specialized community of practitioners focused on neural network architectures which can contain tens or hundreds of thousands of nodes connected in a graph structure. Each node has a weighting learned by presenting the architecture with labeled training examples (Lundervold and Lundervold, 2019). Deep learning occurs when algorithms propagate back and forth through the network comparing labeled information, like a handwritten number or an image of a dog, and tuning the weights with high performance computing assets to produce impressive results in areas like computer vision, speech recognition, and natural language processing.

It is important to note that in machine learning and deep learning, the machine only learns (and therefore, knows) what it is taught and results do not extrapolate outside of what it is taught. Current solutions in this space represent artificial narrow intelligence, not artificial general intelligence. Additionally, deep learning only exceeds the power of the more general machine learning approaches when data is abundant, as in millions of observations as opposed to tens or hundreds of thousands of data samples. Having set the scene, the remainder of the section looks at the fundamental of taking a mindful approach to modeling in AI.

6.3.1 Mindful modeling approaches: a mark of thoughtful work

Data science projects begin with questions and end in answers. Providing relevant answers demands a thoughtful exploration of the question space, especially while challenging assumptions and considering sources of potential bias. Gendron and Killian (2020) provide the **Q *{d m}* A Framework** describing four key functions of data science while communicating the relative importance of each function.

The **Q**uestion and **A**nswer functions are larger than ***d***ata and ***m***odel functions to emphasize the importance of understanding the question before moving into technical activities. Why? Because many practitioners focus on a modeling technique (what model to build) without allowing it to emerge from the nature of the question. The data and model functions are depicted in lowercase and brackets ***{d m}*** to communicate their behind-the-scenes nature, relative to the question and answer. Finally, the ordering of data then model reminds practitioners that the model form follows the data used (Gendron and Killian, 2020). Improvements to that original framework are presented in Fig. 6.2.

The additional **M** (for Mindful Modeling) appears coequal with **Question** and before **data** to focus on building and scrutinizing a mental model before incurring the expenses of collecting, preparing, and modeling data.[1] Mindful modeling could fill an entire chapter. Page (2021) introduces his concept called model thinking by highlighting the role that models, and the thinking they generate, have benefited society through time. Page notes the complexity of the modern world with its diversity in thought and global interactivity combine in many and surprising ways, like the COVID-19 pandemic. Despite an abundance of data (or perhaps because of them) one way to make sense of uncertainty is through models.

> *Models are formal structures represented in mathematics and diagrams that helps us to understand the world. Mastery of models improves your*

[1] IBM's *Foundation Methodology for Data Science* locates two action stages, "Business understanding" and "Analytic approach", before "Data requirements" (Rollins, 2015).

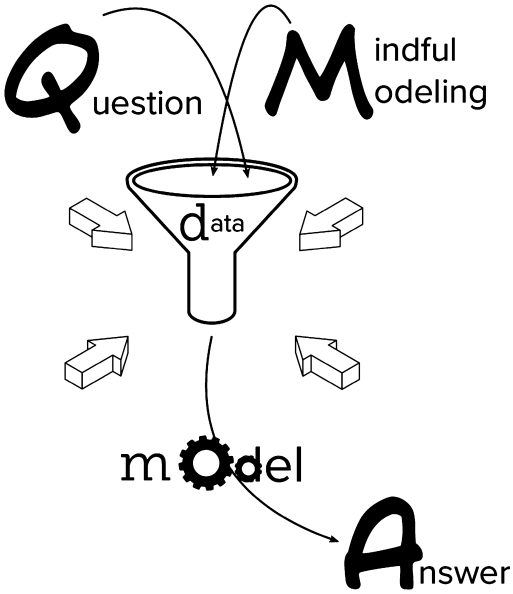


FIGURE 6.2 The **Q M *{d m}* A framework** — Enhancements on the Gendron and Killian 2020 framework to incorporate mindful modeling as a coequal function prior to data collection.

> *ability to reason, explain, design, communicate, act, predict, and explore (Page, 2021).*

We continue the discussion of modeling by sharing this well-worn advice from our professional experiences.

Start a project with small, understandable, and auditable models — then add to them.

6.3.2 Start **S.M.A.L.L. (Specific-Mindful-Attainable-Limited-Lucid)**

Starting small is not glib advice; rather, it is a sincere commentary based on the state of modeling witnessed in our experiences. Technological advances in algorithms, computational power, and cloud-based platform availability encourage a process of going straight from big data to big complexity (and possibly a big mess) when modeling and communicating results to a

business. The mnemonic **S.M.A.L.L** provides guidance when launching a modeling effort.

- **S**pecific — working on one problem unearths other related problems. Keep the initial modeling focused on a particular subregion of the modeling space (conceptual modeling, Section 6.3.2.1)
- **M**indful — captured in the discussion on mindful modeling (Section 6.3.1 and the **Q M*{d m}* A framework**. There are practical approaches to help achieve this state of mind (conceptual modeling, Section 6.3.2.1; group model building, Section 6.3.2.2)
- **A**ttainable — improve time spent modeling by assessing the attainability of a conceptual model. Techniques include testable implications of causal structures (causal modeling, Section 6.3.2.3)
- **L**imited — real-world problems are messy and complex. Continue to scope down the problem in the initial stages to allow for technical audits through the modeling approach (conceptual modeling, Section 6.3.2.1)
- **L**ucid — explainability has become a critical aspect of modeling to address unintended bias in AI systems. Establish clear thoughts on where human interaction is (and is not) desired in the modeling process (causal modeling, Section 6.3.2.3)

Each element of starting S.M.A.L.L. is presented as a subsection below.

6.3.2.1 Conceptual modeling

Conceptual models are used in a wide array of domains from education to science. We propose a working definition:

> *A **conceptual model** explains a real world system in a single image along with accompanying simplifications, limitations, and assumptions.*

Consider the water cycle shown in Fig. 6.3. This conceptual model (water cycle) possesses the key characteristics noted in the definition:

- Single image: like a good resume, limit the amount of material for a conceptual model to a single page

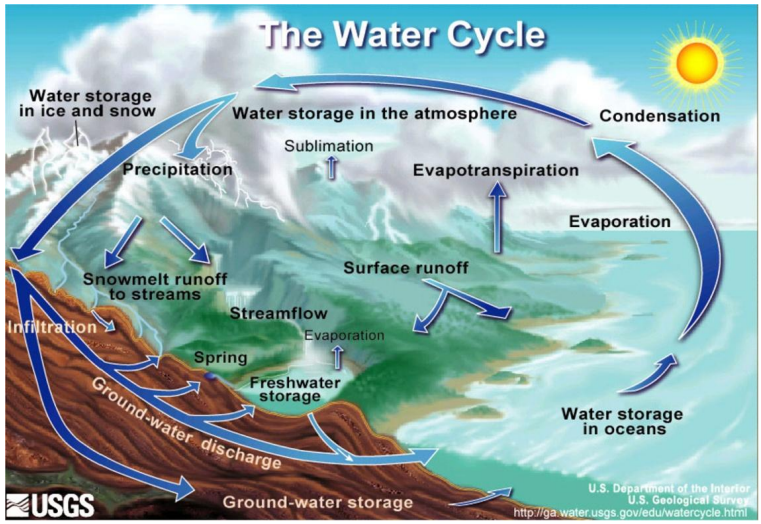


FIGURE 6.3 The water cycle: An example of a conceptual model to simplify the complex actions of atmospheric science into evaporation, condensation, precipitation, and collection (Even, 2021).

- Simplifications: the rain cycle teaches essential atmospheric science principles to children. How is this possible? By focusing on the critical aspects necessary to capture the phenomenon of interest
- Limitations: communicating those aspects the model will *not* address[2]
- Assumptions: "a specific supposition of the operational environment that is assumed to be true, in the absence of positive proof, essential for the continuation of planning" (United States Department of Defense, 2021)

Each key characteristic of a conceptual model helps with a mindful modeling approach encouraging diverse thinking about the question before collecting data. The overall value of a conceptual model is to provide insight and drive the direction of the computational modeling. That said, creating conceptual models can be challenging for a number of reasons, including differing experience bases among business and technical teams, learning when and how to simplify a real-world problem for conceptual modeling,

[2] One natural limiter may be found in the notion of documentation debt (Bender et al., 2021), where the exercise of documentation encourages accountability of the model; size does not imply diversity.

and capturing a concept in a timely manner (Luna-Reyes et al., 2007). Outside of the field of data science, conceptual models that lend themselves to quantitative measurement (e.g., health belief model), drive the modeling work of many social science researchers. Social scientists recognize an important part of modeling is having a strong conceptual familiarity with theoretical and empirical literature in a particular research area (Kline, 2011). Historical perspective from the field of systems dynamics can be helpful in developing solutions to the challenges of conceptual modeling.

6.3.2.2 Group model building process

Systems dynamics is a methodology to generate models of complex systems as a means of understanding interactions (Bala et al., 2017). The first computational form of this modeling occurred in 1958 when Richard Bennet developed SIMPLE (simulation of industrial management problems with lots of equations), opening the door to working with complex systems that might not otherwise be captured in straightforward differential equations (System Dynamics Society, 2021). Over time, a technique called **causal loop diagramming** emerged for gathering causal mechanisms underlying the model. Causal loop diagrams serve as a type of conceptual model to show the relationships among measurable elements and the interventions that would increase or decrease their levels.

Conceptual modeling, to include causal loop diagramming, capture better representations of the real world, so long as domain and technical expertise come together to generate the conceptual model. Luna-Reyes et al. published an article with techniques for a team approach to this process called **group model building** (2007). Their group model building approach contains nine scripts to run the full process. Script 4 focuses on the conceptual model, and Fig. 6.4 provides a representative outcome of that script.

At its core, *Script 4: concept model* fosters interdisciplinary collaboration of a workshop team with a common goal of creating a conceptual model:

1. Users and experts join in dialogue to conceptualize the mechanics of the phenomenon of interest based on working knowledge
2. Modelers encode these initial results (causal loop diagrams) and run output traces using those conceptual models

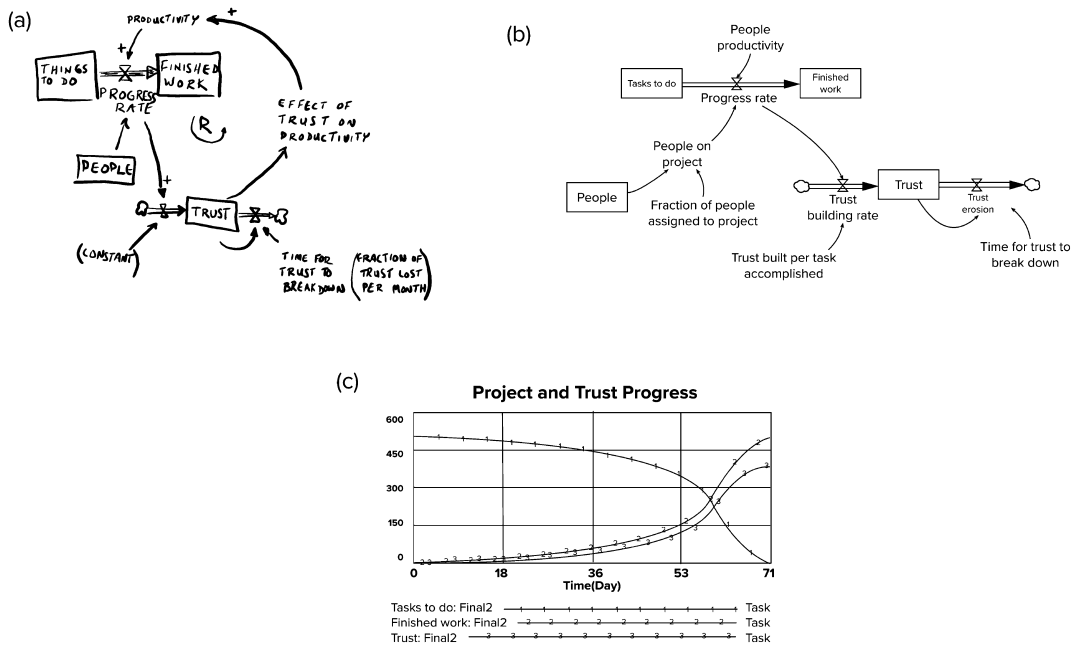


FIGURE 6.4 Group model building results: A technique used in systems dynamics modeling to gather domain experts and capture a conceptual model (panel a) that modelers use to expedite prototyping (panel b) and assess the calculated outputs (panel c) against domain knowledge and historical data (Luna-Reyes et al., 2007).

3. Users and experts review the output from their instincts and improve the first model to capture missing relationships and eliminate undesired qualities (potential bias)

The scripts have pedagogical purposes in mind. Script 4 contains a summary of the objectives, process, and assessment. Recognizing the challenges of concept modeling, the script also offers heuristics to aid in their development (Luna-Reyes et al., 2007). Group model building supports S.M.A.L.L. by gathering multiple perspectives on a problem prior to more intense modeling work later in the process. Related to this is the idea of causal modeling.

6.3.2.3 Causal modeling

Judea Pearl, a professor of Computer Science at the University of California, Los Angeles, has worked in the field of AI for over 40 years. He pio-

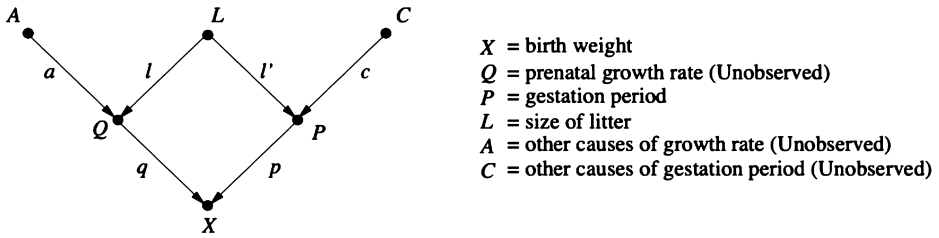


FIGURE 6.5 Causal diagram: The essential object in causal modeling that captures a cause-effect relationship, which are later assessed using tests of conditional independence implied by the structure (Pearl and Mackenzie, 2018).

neered practical methods for including causation in AI. His seminal work in Bayesian networks sought to express the mathematical correlations in a model so that casual expressions would emerge; however, his work identified the limitations of Bayesian networks for causality. Those limitations resulted in his invention of a *do* calculus based on causal diagrams. Fig. 6.5 provides an example of a causal diagram (Pearl and Mackenzie, 2018).

The value of this technique, even without the associated *do* calculus, is to focus thinking on the drivers of an effect. As Pearl notes "think of causation as a form of listening; X is a cause of Y if Y listens to X and decides its value in response to what it hears" (Pearl et al., 2016). Even in cases where there is not an absolute cause and effect relationship, we have found utility in discussing the notion of causality with clients by sketching out causal diagrams, beginning with a final effect and branching backwards to possible causes. Surprising causal elements are often identified in the discussion. This, in turn, focuses model developers on data collection needs, whether from internal sources or by acquiring required data external to the organization.

6.3.3 Inference in modeling

"Statistics is the science of learning from experience" (Efron and Hastie, 2016) and inference is the process where that learning occurs. We observe small pieces of the world and hypothesize generalizations from observed behavior. Galileo's observations and precise measurements of planets, pendula, and falling objects paved the way for Newton's unifying and universal laws of motion, expressed in compact mathematical notation. Statistical

inference is our prime tool for linking observed data to the mathematics defining a more general and universal theory.

As Efron and Hastie (2016) point out, there is an algorithmic component to inference that needs to be separated from the actual act of generalizing from observation. If we think of observation and data as the raw materials for building a theory, statistical algorithms are the tools we use to shape and combine those materials with the goal of building an edifice or structure that can stand up to the highest levels of scrutiny. These algorithms can be as simple as taking an average of a set of measurements or as complicated as the process of building a neural net that can play chess or diagnose a disease. In both cases, they are tools for making a generalization from a particular set of observations and measurements. Of course, a set of measurements can be consistent with many different possible theories. The Ptolemaic geocentric model of the solar system explained the motion of the planets, so also does the heliocentric model of Copernicus as modified by Kepler and Newton. How do we know which theory to choose when making an inference? This section examines statistical inference from both the frequentist and probabilistic points of view and causal inference.[3]

6.3.3.1 Frequentist (Fisherian) inference

Modern frequentist inference, developed by Galton, Fisher, and Pearson, produced the earliest tools for attacking problems related to empirical observation. For these scientists, the act of observing and measuring was error-prone and algorithms were needed not just to calculate quantities derived from the data (e.g., the average or standard deviation), but also the accuracy of the prescribed algorithm. For example: to estimate the average age of men in New York or London, a survey of a smaller sub-population might be taken and the broader population estimate inferred from the sample estimate. Depending on how the sub-population was selected, there may be more or less error in the sample.

[3] The concepts presented in the treatment of this fundamental area are not intended to provoke arguments about the relative merits of one method over another. The intention of this subsection is to increase awareness of techniques that continue to emerge based on the calls for causal inference from revered practitioners of AI.

Nonetheless, the standard error (a quantity calculated and defined in terms of the sample data) estimates the accuracy of the sample average. By estimating the errors in our calculated quantities, we assess how well the measured data support a given hypothesis or theory. Note that the same data used to develop the statistic in question (here the average) are also used to estimate the error in the statistic. Unfortunately, as many first-year statistics students have learned, this approach does not always help with distinguishing between similar theories that are supported by the data[4] (i.e., have similar or overlapping uncertainties). Moreover, because the hypothesis is seen as fixed and the data variable, the Fisherian approach can lead to the practice of p-hacking, where an analyst cherry-picks the data that best fits the desired hypothesis by performing an experiment multiple times and explains away or ignores the rest of the data, often comprising a significant fraction of the total collected data. An alternative approach to statistical inference is provided by Bayesian reasoning, where the data are seen as fixed and the goal is to find the hypothesis that best fits the data. Chen notes many challenges in statistical predictive models (Chen et al., 2021). This is the subject of the next subsection.

6.3.3.2 Probabilistic (Bayesian) inference: a gateway to causal inference

Bayesian methods have a colorful history. Controversy surrounded this method since its development in the late 18th century, predating Fisherian approaches. The methods have made continued contributions to the field of AI and gained more favorable viewpoints in the late 20th century (Mcgrayne, 2011). Readers interested in a more comprehensive treatment of the dramatic history of these methods may consider reading McGrayne's book (see Further Reading at Section 6.6).

Bayes' theorem (and more recently, Bayes–Price theorem) was developed by the Reverend Thomas Bayes (c. 1701–1761) and published by his friend, Richard Price, in 1763. At around the same time in Europe, Pierre-Simon

[4] Undergraduates are taught a collection of statistical recipes, rather than an appreciation for modeling. Though statistical inference works in many circumstances, the inexperienced analyst can apply frequentist recipes in a manner that can lead to misunderstandings or unintentional abuse of the statistics.

LaPlace was developing the roots of Bayes' theorem.[5] What began as a technique to explain probability in games of chance, the theorem would become the way to interpret the inverse probability, the probability of a stated hypothesis given (conditioned on) the observed data. As noted earlier, Bayes' theorem differs from Fisherian methods in that Bayesian methods treat the data as fixed and the parameter describing the probability distribution of the hypothesis as a parameter and measures its uncertainty. Frequentist (Fisherian) methods are most effective with normal distributions, because the residuals (error terms) are also normally distributed. This is not true of most other distributions. It is worth noting that Bayesian and frequentist estimates converge when data is abundant.

One other highlight of Bayesian methods is that their output includes not only the posterior probability, but also a credible interval. These can be much more intuitive to customers, because they indicate the chance of a hypothesis occurring as a probability. This is quite different from the difficult-to-describe confidence intervals resulting from frequentist methods. It is not the purpose of this subsection to argue that Bayesian methods are better than Fisherian methods. Rather, the purpose has been to raise awareness of an approach that is not often taught in core education. It turns out that Bayesian networks are influential: Bayesian methods lie at the heart of Bayesian networks, otherwise referred to as Bayes belief nets. Judea Pearl developed these graph structures to study causality, which is picked up in the next section.

6.3.3.3 Causal inference: tempting the trope that "correlation does not imply causation"

The back cover of Pearl and Mackenzie's *The book of why: The new science of cause and effect* reads

> *"Correlation is not causation." This mantra once led to a virtual prohibition on causal talk. Today that taboo is dead. The causal revolution, instigated by Judea Pearl and his colleagues, has cut through a century*

[5] Interestingly, LaPlace's work constitutes the contemporary form of the theorem today, but it is still attributed to Bayes (Bayes–Price).

> *of confusion and established the study of cause and effect on a firm scientific basis (Pearl and Mackenzie, 2018).*

Readers familiar with Pearl's work may recognize the underlying message of the quotation above. Similarly, Woodward notes "causal inference is concerned with a very specific kind of prediction problem: predicting the results of an action, manipulation, or intervention" (Woodward, 2005). The literature reveals an opinion of modern scholars that the trope of "correlation is not causation" has led to roadblocks in the progression of AI. This subsection presents a quick overview of the underlying ideas of causal inference as they relate to AI and AI assurance. Consider Fig. 6.6. The ladder of causation (Pearl and Mackenzie, 2018) is a summary (a conceptual model) of the seminal work of Pearl starting in the 1980s that developed out of his efforts to model causality with Bayesian networks. The figure is whimsical, yes, yet poignant in its message. The caption as found in the source reads:

> *The Ladder of Causation, with representative organisms at each level. Most animals as well as present-day learning machines are on the first rung, learning from association. Tool users, such as early humans, are on the second rung, if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumeably this is how babies acquire much of their causal knowledge. On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena (Pearl and Mackenzie, 2018).*

The first rung of the ladder depicts machine learning (a subset of AI) and focuses on the activities seeing and observing. It answers the questions *"what if I see…?"* and *"how are the variables related?"* This is correlation. For instance, wind speed and the movement of leaves may have strong correlation in a data set, but will not provide any information on whether the moving leaves caused the wind or if the wind caused the leaves to move. In a business setting, many challenges demand explanation at a higher level of causation, such as the second rung of the ladder, intervention.

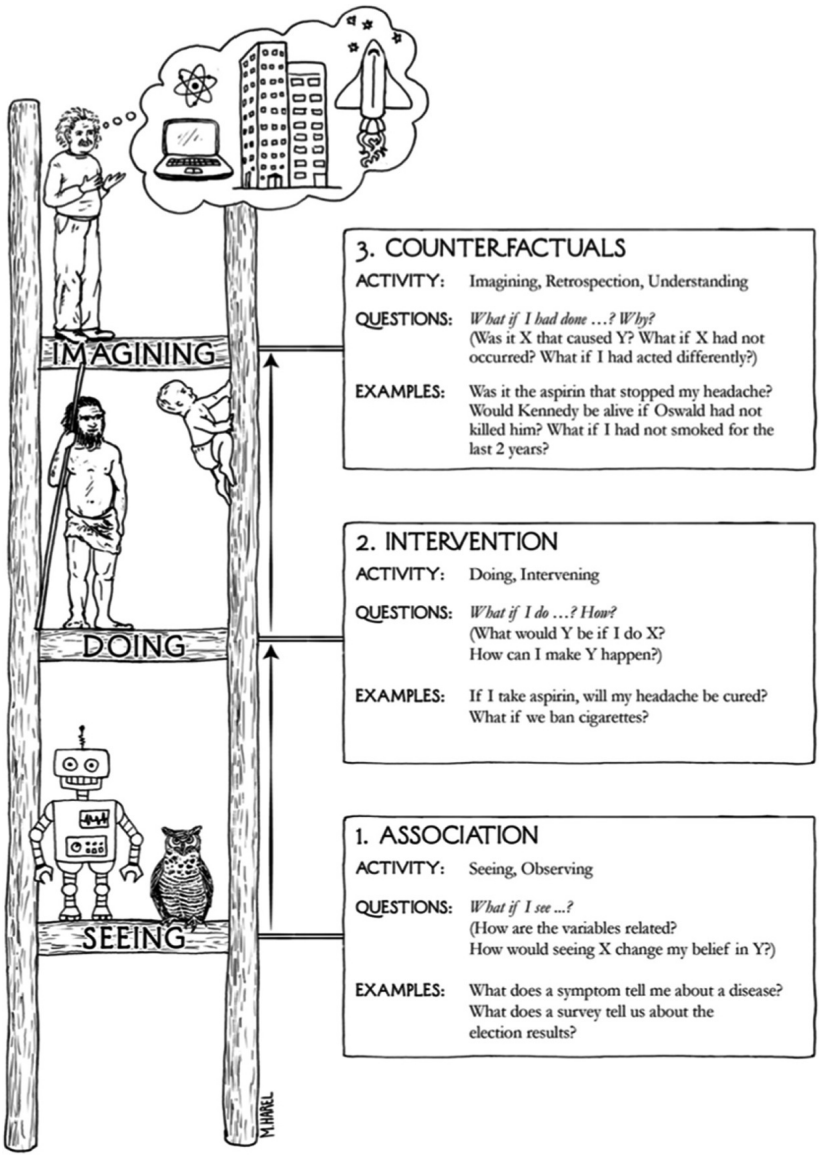


FIGURE 6.6 The ladder of causation: A conceptual model showing the relationships among increasing levels of causal modeling punctuated with generalized questions each level addresses (Pearl and Mackenzie, 2018).

At the second level, the focus of the activity shifts to doing or intervening and addressing the questions *"what if I do. . . ?"* and *"what would Y be if I do X?"* Business decision-makers and leaders are often interested in the

drivers of an end state. They are aware of the effects, and without using the word cause, they are looking for drivers (causes), where they can intervene by taking actions. Predictive and prescriptive analytics are not mere games. Consider the plight of a hardware store owner who must predict the volume of snowblowers that will sell and (most importantly) place a snowblower order in the spring. Ultimately, that store owner is compelled to take action (intervene) by making a decision and committing to an order size. What will cause the expected volume to increase or decrease compared to last winter? Similarly, a healthcare provider may benefit from a tool that aids with the cognitive strain brought on by the effortful System 2 thinking related to determining which one of her patients is at risk for sepsis (an infection of the blood stream resulting in a cluster of symptoms, such as drop in a blood pressure, increase in heart rate, and fever).

The idea of causal inference remains a sleeping giant within the AI community. Schölkopf et al. support this in their article calling for more causal work, noting that if more work is not done to implement causation it could dampen progress in AI (2021). Meanwhile, Gelman and Vehtari (2021) identify causal inference as one of the eight biggest statistical ideas of the last 50 years. This is borne of the trend beginning in the 1980's with Pearl's work in AI developing Bayesian networks, leading to the need for a mathematical treatment of causality. By the first decade of 21st century, one finds Pearl's seminal works published on causality. Adding to the body of knowledge are a number of articles (Fernández-Loría and Provost, 2021; Hünermund and Bareinboim, 2019; Shrier and Platt, 2008) and the *Journal of causal inference* (Imai et al., 2021). In addition, one finds work on causal inference within professional conferences, such as the European causal inference meeting and the American causal inference conference, the web-based platform for knowledge exchange (causalscience.org, 2021), and a python library called "DoWhy," which is an open-source library developed by Microsoft research for end-to-end causal inference (Sharma and Kiciman, 2020). Doctoral candidates working in causal inference have matriculated to post doctoral work and are making these techniques more available.

Moreover, social science researchers pioneered the statistical technique called structural equation modeling (SEM) out of a need to model hypoth-

esized relationships among observed (manifest) and unobserved (latent) variables (e.g., race and intent to engage in social action). As part of this important latent variable work, social scientists also learned how to apply invariance testing as a way to assure that prediction models work across various categories of data (e.g., minority populations versus majority) (Prosperi et al., 2020). In addition, path diagrams (visual representations of the hypothesized associations and dependencies) are the observed versions of latent variable SEM models and are critical when studying causality. It is through those path models (such as SEM) that the AI community can visualize causal inference. This bodes well for the future. It is our opinion that although industry executives will continue to appreciate prediction, they prefer to learn about "levers" they can act upon. In total, these developments will continue to tempt the trope that correlation is not causation.

6.4 Bolstering AI assurance: reducing biases with inferential methods

All politics is local.

— US Rep. Thomas "Tip" O'Neill, Former Speaker of the US House of Representatives (1977–1987)

AI has much to offer humanity. Yet, it can suffer from unintended biases and result in unethical outcomes if left unchecked. A contributing factor to this technological tension is the differing adoption rates between the technology itself and its governance. It is not uncommon for "prevention initiatives" to lag attention on "crisis management." This is often seen in politics, environmental challenges, and even how we choose to care for ourselves; will it be a visit to the gym or the cardiologist? Within the realm of AI, a balancing force is found in AI assurance.

Similar to politics, one could argue that all AI is local, meaning that despite well-adopted techniques, an appropriate use of AI requires an appreciation of the problem context, technique, audience, and intended use. This section contemplates AI assurance being mindful of the fundamentals from earlier in this chapter: bias, modeling, and inference. A working scenario is

used to frame the thought process, but we begin with a brief exploration of AI assurance.

6.4.1 What is AI assurance?

There is no standardized meaning for AI assurance. Batarseh et al.[6] recognize this gap in practice, where practitioners must rely on existing industrial techniques, such as verification, validation, and testing to compensate for the lack of specificity. They propose a definition of **AI assurance**:

> *A process that is applied at all stages of the AI engineering life cycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users (Batarseh et al., 2021).*

They also propose five considerations when putting AI assurance methods into practice, which are summarized in Table 6.2.

6.4.1.1 Working scenario: mitigating bias in healthcare through AI assurance

The COVID-19 pandemic illuminated the fact that racism in healthcare leads to health disparities. Healthcare experts often use patient race as the predictor variable in their models to study health disparities; however, the biological theory of race was debunked over 30 years ago. The biological determinists who argued about the genetic inferiority of non-White races have failed to provide evidence to support this assumption. In contrast, evidence continues to mount that health is determined by social factors, such as racism, not by race (Gould et al., 1996). Risks caused by the current discourse between healthcare leaders and providers about the actual predictors of poor patient outcomes (i.e., it is social determinants of health brought on by institutional racism, and not a person's race) can be miti-

[6] Batarseh and Freeman are Virginia Tech faculty members of the Commonwealth cyber initiative, which notes "it is imperative that trust and assurance mechanisms are baked into the development and deployment process. AI systems must be deemed reliable, explainable, unbiased/fair, and privacy-preserving" (Commonwealth Cyber Initiative, 2021).

Table 6.2 AI Assurance — Five considerations for defining and implementing assurance methods (Batarseh et al., 2021).

Consideration	Description
(1) Data quality	with outcomes in mind, verify data elements are free from issues that could hurt assurance in production
(2) Specificity	focus assurance methods to the intended use of the model and data
(3) Addressing invisible issues	include assurance methods proactively in AI development procedures, rather than attend to crisis management when visible issues emerge
(4) Automated assurance	implement a form of human-NOT-in-the-loop to mitigate the risk of human interference in evaluating for bias
(5) The user	involve the user community in "expert-relevant (non-engineering) domains such as healthcare, education, economics, and other areas" to get better insights to subjective explainability matters

gated by AI assurance. Indeed, we argue that the five considerations of AI assurance as outlined in Table 6.2 can counter bias. Section 6.4.2 walks through a mindful approach to applying AI assurance design to the working scenario.

6.4.2 Contemporary AI: mindful modeling before data engineering helps reduce bias

Mindful modeling based on one's sense of the operational environment not only helps to extract away complexities, but also provides a reference to compare early modeling work. As models continue to focus on more complex issues, it will become more difficult to defend the output as unbiased and of the highest integrity. From where will the complexity emerge? Consider where AI has yielded its more popular contributions since the last AI winter:

- computer vision: despite the challenges of "seeing" a cat or a dog within an image, it is a deterministic problem supported by physical realities
- speech recognition and natural language processing: despite the many languages, dialects, and individual stylings, it possesses an underlying (quasi-deterministic) structure

- natural sciences: despite many stochastic processes, it enjoys theories and accepted truths, such as gravitational pull, pressure differential of air masses, and many other physics-based realities

The problem sets characterizing accomplishments since the last AI winter are deterministic and structural in nature; they are rational problem sets. Missing from the list above is human behavior. As Daniel Kahneman notes, it makes very little sense to think about people in terms of rationality stating, "I never use the word irrational" (University of New South Wales, 2021). In that conversation Kahneman went further by saying that the meaning of rationality is a technical construct of decision theory, but impractical for human minds and decision-making.

There are few (and sometimes no) deterministic anchors from which we can predict human behavior. Volatile global stock market indices are a testament to predictions that cannot be made. Thankfully, a great many problem sets lack life-threatening impacts from poor predictions, but some problem sets do fall in the category of life-threatening and others can be life changing (sometimes for the worse). Applications run through machine learning and AI systems for mortgages, schools, employment, and housing can have life-changing consequences.

In this section, we contemplate the public health crises resulting, in part, from unknown (and undesired) bias inherent in health care systems and practitioners. Combining the broader knowledge about S.M.A.L.L. approaches (Section 6.3.2) with the working constructs for AI assurance (Section 6.4.1) provides the start of practical applications for the working scenario. Table 6.3 presents principled guardrail questions detailed below.

6.4.2.1 Question 1: what is the basis of ground truth for teaching the machine?

Consideration: mindful of data quality (1) and the user (5) [numbers reference elements in Table 6.2]

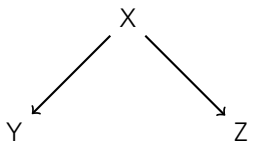
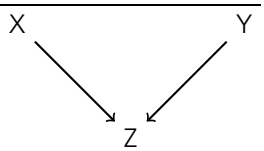
Strategy: causal modeling (in the spirit of **A**ttainable) [Section 6.3.2]

Causal inference makes use of causal diagrams also known as directed acyclic graphs (DAG) to show the relationship from a cause to an effect. These relationships are elicited from domain experts. An interesting as-

Table 6.3 Principled and Practical AI Assurance — General questions for AI assurance design.

Question	**Consideration**	**S.M.A.L.L. Strategy [Section 6.3.2]**
1) What is the basis of ground truth for teaching the machine?	data quality (1) and user (5)	Employ causal modeling *in the spirit of* ***A****ttainable*
2) Who determines when predictive analytics are used in decision-making?	specificity (2) and automated assurance (4)	Employ conceptual modeling *. . . in the spirit of* ***S****pecific* Employ causal modeling *. . . in the spirit of* ***L****ucid*
3) When is a problem cognitively complex enough to obscure bias present in decision making?	invisible issues (3) and user (5)	Employ conceptual modeling *. . . in the spirit of* ***L****imited* with group model building *. . . in the spirit of* ***M****indful*

Table 6.4 The three types of DAG elements: All causal inference diagrams are composed of these three elements.

Name	**Notation**	**Structure**
Chain	$X \rightarrow Y \rightarrow Z$	
Fork	$Y \leftarrow X \rightarrow Z$	
Collider	$X \rightarrow Z \leftarrow Y$	

pect of DAGs is their ability to capture a type of conceptual model, while also generating testable implications (Pearl et al., 2016). Table 6.4 depicts a sketch of the three types of causal diagram DAG elements along with their notation and structure.

Each of the elements in the table include conditional independence that would appear in data if the causal influence diagram reflects the real world.

Moreover, they are testable with data. Any elements that do not pass the testable implication must be re-examined for structural appropriateness. Eliciting the structure before modeling the data is a form of "mindful modeling." This aspect of causal inference is a key benefit of the technique for AI assurance. It can elicit structure which supports conceptual modeling, and it sheds light on hypotheses that may be conducted in light of frequentist inference.

6.4.2.2 Question 2: who determines when predictive analytics are used in decision-making?

Consideration: mindful of specificity (2) and automated assurance (4) [numbers reference elements in Table 6.2]

Strategy: employ conceptual modeling (in the spirit of **S**pecific) and causal modeling (in the spirit of **L**ucid) [Section 6.3.2]

Non-observable traits of the AI model (structure, latent variables, and indiscernible mathematical formulations) are challenges (Hyttinen et al., 2015; Zhao and Hastie, 2021); however, the use of Bayesian methods can offer benefits.

"Simpson's Paradox refers to a phenomena whereby the association between a pair of variables (X, Y) reverses sign upon conditioning of a third variable, Z, regardless of the value taken Z." The paradox is observing inverse correlations when comparing the output from an entire population versus sub-segments of the population (Pearl, 2014). This has caused much confusion in studies and in real-life trials for decades. Fig. 6.7 shows one example of Simpson's paradox.

The figure shows the two-way relationship of cholesterol levels based on exercise. Without controlling for other factors (such as age), the visualization presents a startling finding: cholesterol levels **increase** with higher levels of exercise! If that seems counter intuitive, then you are not alone. Pearl notes, "To resolve this problem, we once again turn to the story behind the data" (2014). The factor *Age* is causal to both the treatment (more exercise as people age) and the outcome (higher cholesterol as people age).

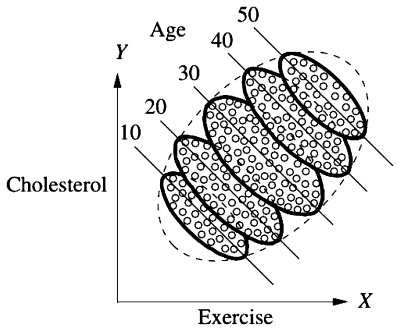


FIGURE 6.7 Simpson's paradox: A two-dimensional data relationship from a medical study showing how the population correlation (positive) is reversed when looking at subpopulations controlled by age (negative) (Pearl et al., 2016).

6.4.2.3 Question 3: when is a problem cognitively complex enough to obscure bias present in decision-making?

Consideration: mindful of invisible issues (3) and the user (5) [numbers reference elements in Table 6.2]

Strategy: Employ conceptual modeling (in the spirit of **L**imited) with group model building (in the spirit of **M**indful) [Section 6.3.2]

Capturing a group's view of reality in a conceptual model allows it to be communicated then validated. All this occurs before creating the computational model. One potential benefit is that a conceptual model could help identify potential biases in thinking that might be obscured by, and not visible in, the data or the modeling approach. Meanwhile, group model building (Section 6.4) provides ways to elicit group ideas in a structured approach and provide domain expert inputs on prior beliefs (in the use of probabilistic inference) and causal diagrams (in the use of causal inference).

6.4.3 Considering the level of system predictability when designing AI assurance

In closing this contemplation of applied AI assurance, it is worthwhile considering if there are circumstances where AI proposed for a problem requires modeling beyond the deterministic solution space common in many AI approaches. Where is it prudent to be satisfied with unbiased under-

standing rather than predictability? FAANG companies[7] exemplify this by making recommendations on movies, books, and entertainment, rather than predictions. Barnes et al. reflect on the nature of predictability based on their work with atmospheric science, a field that yearns for better and longer term predictions, while realizing a foolhardy prediction of the unpredictable can come at the price of human life.

Their work focuses on using AI to understand if a problem set could benefit (is predictable) from AI (Barnes et al., 2020). This perspective challenges the wisdom set forth by William Shakespeare's character Brutus from the play Julius Caesar who replied, "No, Cassius, for the eye sees not itself but by reflection, by some other things" (Shakespeare et al., 1974). Perhaps, A-"eye" can see itself?

Barnes et al. (2020) use neural networks to find climactic conditions that are indicators for increased predictability: comparing neural network insights to understood theory of natural science and physics. AI is a collection of techniques that learn complex and well-hidden patterns within information. Generalizing this approach in atmospherics highlights there are degrees of AI-solvable problems, ranging from deterministic to stochastic to intractable and from empirical to experimental to theoretical.

Understanding the limitations of today's predictive approaches allows developers to also understand the associated impact to traditional validation techniques for less-predictable problems. Traditionally, the role of model governance relies on validation techniques, such as back-testing (assessing model performance with out-of-sample data) and operational monitoring (detecting model drift over time). The discipline of model validation is captured in the literature throughout the decades (Anderson and Woessner, 1992; Landry et al., 1983; McLean et al., 2012; Sargent, 2011) and is an established technique to support AI assurance; however, validation techniques may suffer from inherent or unidentified bias because of the human role in validation design. The techniques above rely on a) backtesting against historical outcomes, that could themselves carry with them a

[7] FAANG is an acronym representing the stocks of American technology companies: Facebook, Amazon, Apple, Netflix, and Alphabet (Google).

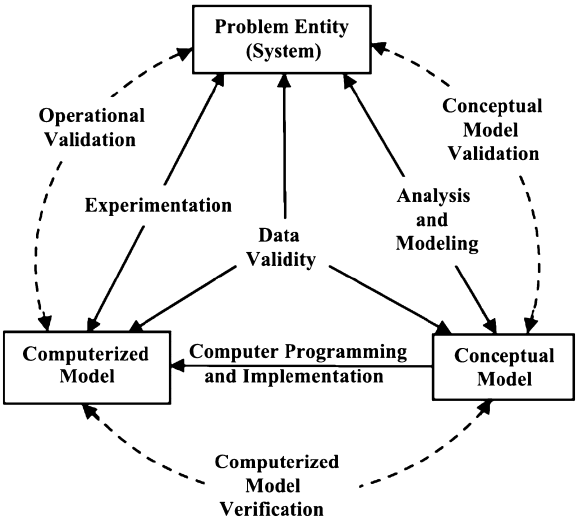


FIGURE 6.8 Model validation as elements: Model validation is a complex system of related paradigms of validation, each relevant to the pair of elements considered (Sargent, 2011).

human bias; and b) monitoring against human designed key performance indicators, that may also contain unintentional bias.

Sargent's work on validation techniques provides a useful survey of the validation literature and practical model validation methods that are extensible to AI assurance. Fig. 6.8 provides a simplified presentation of the modeling process. It differentiates among various types of model validation, rather than treating the topic as a unary process.

The figure depicts a total of three broad paradigms (the outer ring) of validation relating the problem with conceptual and computational manifestations of modeling. Two paradigms and three inferential techniques are curated in Table 6.5 as a summary.

6.5 Rest assured: mindful approaches in modeling may help avoid another AI winter

I don't think that any of the human faculties is something inherently inaccessible to computers. I would say that some aspects of humanity are less accessible and creativity of the kind that we appreciate is probably one that is going to be something that's going to take more time to reach. But maybe even more difficult for computers, but also quite important,

Table 6.5 Traditional model validation: The connection of inferential techniques with operational and conceptual model validation.

Paradigm	Relationship	Definition	Inferential Connection
Operational validation	relating the computerized model to the problem	"determining that the model's output behavior has sufficient accuracy for the model's intended purpose over the domain of the model's intended applicability" (Sargent, 2011)	**frequentist** inferential techniques
Conceptual model validation	relating the conceptual model to the problem	"determining that the theories and assumptions underlying the conceptual model are correct and that the model representation of the problem entity is 'reasonable' for the intended purpose of the model" (Sargent, 2011)	**probabilistic** and **causal** inferential techniques

> *will be to understand not just human emotions, but also something a little bit more abstract, which is our sense of what's right and what's wrong.*
> — Yoshua Bengio

AI has and will contribute much knowledge and capability to society. One question is whether another AI winter will set in (Kinsella, 2017) or if practitioners will adapt to avoid this? The noted researcher Yoshua Bengio is public in his call for the increased use of inferential techniques, specifically causal inference, to provide a missing element to current AI approaches. As children we learn right from wrong and effect resulting from cause. These are the building blocks of generalized intelligence. This chapter notes the low-hanging fruit that catapulted this prolonged period of success in AI are disappearing. Structured and physical problem sets, such as computer vision, speech recognition, and natural sciences will not be a satisfying endpoint, merely a way point to the problems surrounding human behavior. Kahneman shared his thoughts to avoid viewing people as rational or irrational, they are human with all the foibles of a system that continues to learn and adapt.

Prosperous societies are based on some type of governance. AI, in its presumed trajectory to be more human-like, will be bound by the same guiding

principles if it is to help society. AI assurance is an emerging field of interest spurred by concerns about the ethical use of AI and desires to develop explainable, unbiased models. As the body of work unfolds for AI assurance, the rich collection of practices from validation and testing can serve as an intermediate foundation for growth. Validation and testing is of human design, and therefore subject to biases. We proposed mindful modeling, a philosophy intertwined with questioning, to avoid premature reliance of data and consider human input to the causal structure of the problem set. An approach called Start S.M.A.L.L. provides practical steps relying on the foundation of psychology and inference to give practitioners a place to begin mindful modeling and target the weaknesses resulting from human bias in the validation design with insights from causal inference to increase assurance in future intelligent systems.

6.6 Further reading

This chapter scratched the surface on large bodies of literature about fundamental subjects of import: bias and causality. The remainder of this book will shed further light on bias within an AI context. Readers interested in fathoming the depths of bias from a psychological context may be interested in these titles:

- *The signal and the noise: Why so many predictions fail—but some don't* by Nate Silver (2012)
- *Noise: A flaw in human judgment* by Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein (2021)
- *Thinking, fast and slow* by Daniel Kahneman (2011)

Recommended reading on the topic of causality and probabilistic modeling include:

- *The book of why: The new science of cause and effect* by Judea Pearl (2018)
- *The theory that would not die: How Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy* by Sharon Bertsch Mcgrayne (2011)
- *Think Bayes: Bayesian statistics in Python* by Allen Downey (2021)

Acknowledgments

Raphael Laufer is a senior Data Scientist and mathematician at Systems Planning and Analysis. We thank him for serving as a technical reviewer on the content relating to inferential methods. He has a PhD in mathematics from the University of California, Berkley and an AB in mathematics from the University of Chicago. In his role, Raphael works within the discipline of data science developing models. He has experience validating models and researching causal inference techniques.

References

Alpaydin, E., 2010. Introduction to Machine Learning. MIT Press.

Anderson, M., 2019. Cincinnati Children's Hospital admits to giving patients improper doses of blood pressure drug. https://www.beckershospitalreview.com/pharmacy/cincinnati-children-s-hospital-admits-to-giving-patients-improper-doses-of-blood-pressure-drug.html.

Anderson, M.P., Woessner, W.W., 1992. The role of the postaudit in model validation. Advances in Water Resources 15 (3), 167–173. https://www.sciencedirect.com/science/article/pii/030917089290021S.

Angwin, J., Larson, J., Kirchner, L., Mattu, S., 2016. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing.

Aronson, B., Keister, L., Moody, J., 2020. Provider bias in prescribing opioid analgesics: An analysis of emergency department electronic medical records. BMC Public Health 21 (1518). https://doi.org/10.1186/s12889-021-11551-9.

Bala, B.K., Fatimah, M.A., Kusairi, M.N., 2017. System Dynamics: Modelling and Simulation. Springer.

Barnes, E.A., Mayer, K., Toms, B., Martin, Z., Gordon, E., 2020. Identifying opportunities for skillful weather prediction with interpretable neural networks. In: NeurIPS 2020 AI for Earth Sciences. https://arxiv.org/abs/2012.07830.

Batarseh, F.A., Freeman, L., Huang, C.-H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 60. https://doi.org/10.1186/s40537-021-00445-7.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623.

Benjamens, S., Dhunnoo, P., Meskó, B., 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. npj Digital Medicine 3 (1), 118. https://doi.org/10.1038/s41746-020-00324-0.

Beran, O., 2018. An attitude towards an artificial soul? Responses to the "Nazi Chatbot". Philosophical Investigations 41 (1), 42–69.

Beresford, B., Sloper, P., 2008. Understanding the Dynamics of Decision-Making and Choice: A Scoping Study of Key Psychological Theories to Inform the Design and Analysis of the Panel Study. Social Policy Research Unit.

Box, G.E.P., 1976. Science and statistics. Journal of the American Statistical Association 71 (356), 791–799. https://doi.org/10.2307/2286841.

Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. PMLR, pp. 77–91.
Butler, A.G., Roberto, M.A., 2018. When cognition interferes with innovation: Overcoming cognitive obstacles to design thinking. Research-Technology Management 61 (4), 45–51.
Cabitza, F., 2019. Biases affecting human decision making in AI-supported second opinion settings. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer, pp. 283–294.
Campbell, F.A., 2001. Inciting legal fictions-disability's date with ontology and the abieist body of the law. Griffith Law Review 10, 42.
causalscience.org, 2021. How to understand and influence behavior using data science. https://causalscience.org.
Cazes, M., Franiatte, N., Delmas, A., André, J., Rodier, M., Kaadoud, I.C., 2021. Evaluation of the sensitivity of cognitive biases in the design of artificial intelligence. In: Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA'21) Plate-Forme Intelligence Artificielle (PFIA'21).
Chen, I.Y., Joshi, S., Ghassemi, M., Ranganath, R., 2021. Probabilistic machine learning for healthcare. In: Annual Reviews of Biomedical Data Science 2021. https://arxiv.org/abs/2009.11087.
Cheshire, W.P., 2017. Loopthink: A limitation of medical artificial intelligence. Ethics & Medicine 33 (1), 7–12.
Chou, J., Murillo, O., Ibars, R., 2017. What the kids' game "telephone" taught Microsoft about biased AI. https://www.fastcompany.com/90146078/what-the-kids-game-telephone-taught-microsoft-about-biased-ai.
Commonwealth Cyber Initiative, 2021. AI assurance. https://cyberinitiative.org/research/ai-assurance.html.
Council of Europe Directorate General Human Rights and Rule of Law, 2019. AFINCoE: Miro Griffiths on AI and technology "rooted in ableism" and social inequalities. Online video. https://youtu.be/sI99ZoE444M.
Culotta, A., Ravi, N., Cutler, J., 2016. Predicting Twitter user demographics using distant supervision from website traffic data. Journal of Artificial Intelligence Research 55, 389–408. https://doi.org/10.1613/jair.4935.
Davenport, T.H., Patil, D.J., 2012. Data scientist: The sexiest job of the 21st century. Harvard Business Review 90 (10), 70–76.
David, R., 2001. Commentary: birthweights and bell curves. International Journal of Epidemiology 30 (6), 1241–1243. https://academic.oup.com/ije/article/30/6/1241/651753.
De Vries, E.N., Ramrattan, M.A., Smorenburg, S.M., Gouma, D.J., Boermeester, M.A., 2008. The incidence and nature of in-hospital adverse events: a systematic review. BMJ Quality & Safety 17 (3), 216–223.
Downey, A.B., 2021. Think Bayes: Bayesian Statistics in Python. O'Reilly Media, Inc.
Efron, B., Hastie, T., 2016. Computer Age Statistical Inference. Cambridge University Press.
Evans, J.S.B.T., 2008. Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology 59 (1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629.

Even, J.M., 2021. Water cycle. Image. https://upload.wikimedia.org/wikipedia/commons/9/94/Water_cycle.png.
Fernández-Loría, C., Provost, F., 2021. Causal decision making and causal effect estimation are not the same...and why it matters. arXiv preprint. arXiv:2104.04103.
Frigg, R., Hartmann, S., 2020. Models in science. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, Spring 2020 edition. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/models-science/.
Garvin, D.A., 2003. Learning in Action: A Guide to Putting the Learning Organization to Work. Harvard Business Review Press.
Gelman, A., Vehtari, A., 2021. What are the most important statistical ideas of the past 50 years? https://arxiv.org/abs/2012.00174v3.
Gendron, J., Killian, D., 2020. Data citizens: Rights and responsibilities in a data republic. In: Bararseh, F.A., Yang, R. (Eds.), Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering. Academic Press, Cambridge, M.A., pp. 9–27.
Gilovich, T., Griffin, D., Kahneman, D., 2002. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press.
Gould, S.J., Gold, S.J., et al., 1996. The Mismeasure of Man. W.W. Norton & Company.
Green, L., Mehr, D.R., 1997. What alters physicians' decisions to admit to the coronary care unit? Journal of Family Practice 45 (3), 219–226.
Haring, K.S., Watanabe, K., Velonaki, M., Tossell, C.C., Finomore, V., 2018. FFAB—the form function attribution bias in human–robot interaction. IEEE Transactions on Cognitive and Developmental Systems 10 (4), 843–851.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY, USA.
Hertwig, R., Pachur, T., 2015. Heuristics, history of. In: International Encyclopedia of the Social & Behavioral Sciences. Elsevier, pp. 829–835.
Hodges, W., 2020. Model theory. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, Winter 2020 edition. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/model-theory/.
Hogarth, R., 2002. Deciding analytically or trusting your intuition? The advantages and disadvantages of analytic and intuitive thought. In: UPF Economics and Business Working Paper. In: UPF Economics and Business Working Paper, vol. 654. https://dx.doi.org/10.2139/ssrn.394920.
Hong, J.-W., Williams, D., 2019. Racism, responsibility and autonomy in HCI: testing perceptions of an AI agent. Computers in Human Behavior 100, 79–84.
Hünermund, P., Bareinboim, E., 2019. Causal inference and data fusion in econometrics. arXiv preprint. arXiv:1912.09104.
Hyttinen, A., Eberhardt, F., Järvisalo, M., 2015. Do-calculus when the true graph is unknown. In: UAI. Citeseer, pp. 395–404.
Imai, Kosuke, Pearl, Judea, Petersen, Maya Liv, van der Laan, Mark J., 2021. Journal of Causal Inference. https://www.degruyter.com/journal/key/JCI/html.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.
Jha, A.K., Larizgoitia, I., Audera-Lopez, C., Prasopa-Plaizier, N., Waters, H., Bates, D.W., 2013. The global burden of unsafe medical care: analytic modelling of observational studies. BMJ Quality & Safety 22 (10), 809–815.

Johnson, R.L., Saha, S., Arbelaez, J.J., Beach, M.C., Cooper, L.A., 2004. Racial and ethnic differences in patient perceptions of bias and cultural competence in health care. Journal of General Internal Medicine 19 (2), 101–110.
Kahneman, D., 2011. Thinking, Fast and Slow. Macmillan, New York.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–291. http://www.jstor.org/stable/1914185.
Kinsella, B., 2017. Gartner hype cycle suggests another AI winter could be near. Blog. https://voicebot.ai/2017/11/05/gartner-hype-cycle-suggests-another-ai-winter-near/.
Kline, R.B., 2011. Principles and Practice of Structural Equation Modeling. Guilford Press, New York.
Krupiy, T.T., 2020. A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective. Computer Law & Security Review 38, 105429.
Kuchenbrandt, D., Eyssel, F., Bobinger, S., Neufeld, M., 2013. When a robot's group membership matters. International Journal of Social Robotics 5 (3), 409–417.
Landry, M., Malouin, J.-L., Oral, M., 1983. Model validation in operations research. European Journal of Operational Research 14 (3), 207–220.
Liao, L., Jiang, J., Ding, Y., Huang, H., Lim, E.-P., 2014. Lifetime lexical variation in social media. Proceedings of the National Conference on Artificial Intelligence 2, 1643–1649.
Lohr, S., 2018. Facial recognition works best if you're a white guy. The New York Times, February 12, B1.
Lord, C.G., Lepper, M.R., Preston, E., 1984. Considering the opposite: a corrective strategy for social judgment. Journal of Personality and Social Psychology 47 (6), 1231.
Luna-Reyes, L.F., Martinez-Moyano, I.J., Pardo, T.A., Cresswell, A.M., Andersen, D.F., Richardson, G.P., 2007. Anatomy of a group model-building intervention: building dynamic theory from case study research. System Dynamics Review 22 (4), 291–320. https://doi.org/10.1002/sdr.349.
Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik 29 (2), 102–127. https://www.sciencedirect.com/science/article/pii/S0939388918301181.
Mcgrayne, S.B., 2011. The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy. Yale University Press. http://www.jstor.org/stable/j.ctt1np76s.
McLean, C., Jain, S., Lee, Y.T., Hutchings, C., et al., 2012. Technical guidance for the specification and development of homeland security simulation applications. https://doi.org/10.6028/NIST.TN.1742.
Mokli, Y., Pfaff, J., Pinto dos Santos, D., Herweh, C., Nagel, S., 2019. Computer-aided imaging analysis in acute ischemic stroke—background and clinical applications. Neurological Research and Practice 1, 23. https://doi.org/10.1186/s42466-019-0028-y.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.
Nagpal, S., Singh, M., Singh, R., Vatsa, M., Ratha, N.K., 2019. Deep learning for face recognition: Pride or prejudiced? arXiv:1904.01219v2.

Nalty, K., 2016. Strategies for confronting unconscious bias. The Colorado Lawyer 45 (5), 45–52.
National Pharmacy Association, 2021. NPA medication safety update Q4 2020. https://www.npa.co.uk/wp-content/uploads/2021/04/NPA-patient-safety-MSO-report-Q4-2020-FINAL.pdf.
Neumann, J.v., Morgenstern, O., 1953. Theory of Games and Economic Behavior. Princeton University Press, Princeton.
Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T., 2013. How old do you think I am?: A study of language and age in Twitter. In: ICWSM. Palo Alto, CA. AAAI Press, pp. 439–448.
Noble, S.U., 2018. Algorithms of Oppression. New York University Press.
Page, S.E., 2021. The model thinker: What you need to know to make data work for you. Basic Books.
Panagioti, M., Khan, K., Keers, R.N., Abuzour, A., Phipps, D., Kontopantelis, E., Bower, P., Campbell, S., Haneef, R., Avery, A.J., et al., 2019. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ 366.
Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: an attentional integration. Human Factors 52 (3), 381–410.
Pearl, J., 2009. Causality. Cambridge University Press.
Pearl, J., 2014. Understanding Simpson's paradox. American Statistician 68, 8–13.
Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal Inference in Statistics: A Primer. John Wiley & Sons.
Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect. Basic Books.
Peersman, C., Daelemans, W., Vaerenbergh, L., 2011. Predicting age and gender in online social networks. In: International Conference on Information and Knowledge Management, Proceedings. Glasgow, Scotland, UK. ACM Press, pp. 37–44.
Phelan, S.M., Dovidio, J.F., Puhl, R.M., Burgess, D.J., Nelson, D.B., Yeazel, M.W., Hardeman, R., Perry, S., van Ryn, M., 2014. Implicit and explicit weight bias in a national sample of 4,732 medical students: the medical student CHANGES study. Obesity 22 (4), 1201–1208.
Prosperi, M., Guo, Y., Sperrin, M., Koopman, J.S., Min, J.S., He, X., Rich, S., Wang, M., Buchan, I.E., Bian, J., 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence 2 (7), 369–375.
Rollins, J.B., 2015. Foundational methodology for data science. https://tdwi.org/~/media/64511A895D86457E964174EDC5C4C7B1.PDF.
Rosales, A., Fernández-Ardèvol, M., 2019. Structural ageism in big data approaches. Nordicom Review 40 (s1), 51–64.
Sargent, R., 2011. Verification and validation of simulation models. In: Engineering Management Review, vol. 37. IEEE, pp. 166–183.
Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y., 2021. Towards causal representation learning. In: Special Issue of Proceedings of the IEEE—Advances in Machine Learning and Deep Neural Networks. https://arxiv.org/abs/2102.11107.

Shakespeare, W., Evans, G.B., Levin, H.S., Harlen, C., 1974. The Riverside Shakespeare. Houghton Mifflin, Boston.
Sharma, A., Kiciman, E., 2020. DoWhy: An end-to-end library for causal inference. Software. https://pypi.org/project/dowhy/.
Shew, A., 2020. Ableism, technoableism, and future AI. IEEE Technology & Society Magazine 39 (1), 40–85.
Shrier, I., Platt, R.W., 2008. Reducing bias through directed acyclic graphs. BMC Medical Research Methodology 8 (1), 70. https://doi.org/10.1186/1471-2288-8-70.
Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychological Bulletin 119 (1), 3–22. https://doi.org/10.1037//0033-2909.119.1.3.
Sterman, J., 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill, Boston.
System Dynamics Society, 2021. Origin of system dynamics. https://systemdynamics.org/origin-of-system-dynamics/.
Todd, K.H., Funk, K.G., Funk, J.P., Bonacci, R., 1996. Clinical significance of reported changes in pain severity. Annals of Emergency Medicine 27 (4), 485–489.
Trainer, T., Taylor, J., Stanton, C., 2020. Choosing the best robot for the job: affinity bias in human-robot interaction. In: Lecture Notes in Computer Science Book Series. In: Lecture Notes in Computer Science, vol. 12483.
United States Department of Defense, 2021. DoD dictionary of military and associated terms. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/dictionary.pdf.
University of New South Wales, 2021. Daniel Kahneman in conversation. Online video. https://youtu.be/I-R3W5GNqxM.
Vedantam, S.H., 2021. The story of stories [audio podcast episode]. In: Hidden Brain. Hidden Brain Media. https://hiddenbrain.org/podcast/the-story-of-stories.
Vincent, J., 2018. Google's Sergey Brin warns of the threat from AI in today's "technology renaissance". Blog. https://www.theverge.com/2018/4/28/17295064/google-ai-threat-sergey-brin-founders-letter-technology-renaissance.
Woodward, J., 2005. Making Things Happen: A Theory of Causal Explanation. Oxford University Press.
Zhao, Q., Hastie, T., 2021. Causal interpretations of black-box models. Journal of Business & Economic Statistics 39 (1), 272–281. https://doi.org/10.1080/07350015.2019.1624293.

This page intentionally left blank

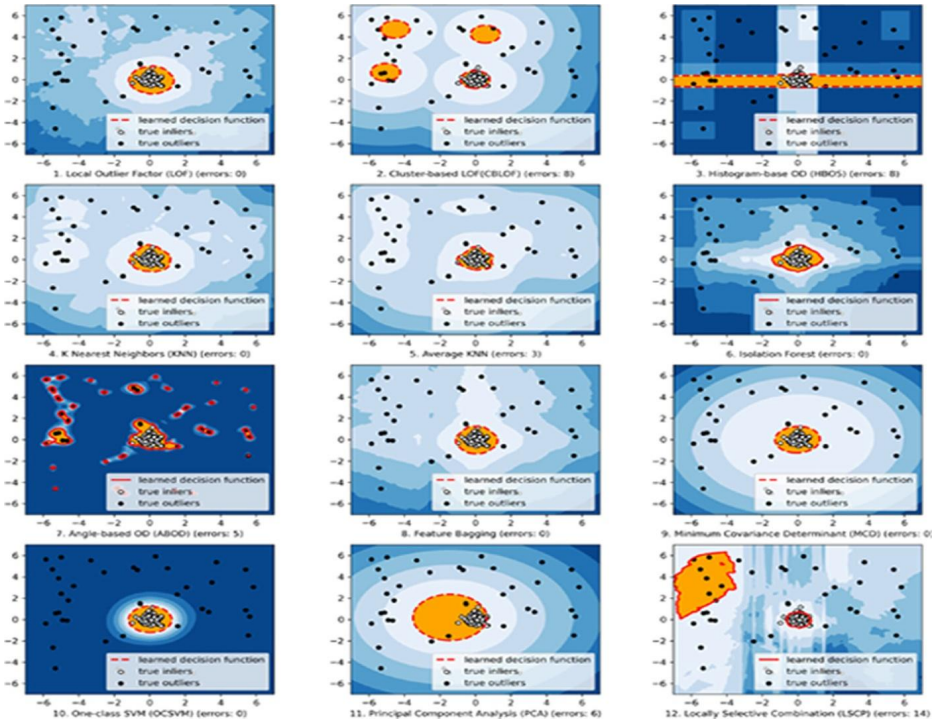
7

Outlier detection using AI: a survey

Md Nazmul Kabir Sikder[a] and Feras A. Batarseh[b]

[a]*Bradley Department of Electrical and Computer Engineering (ECE), Virginia Tech, Arlington, VA, United States* [b]*Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States*

Graphical abstract



Abstract

An outlier is an event or observation that is defined as an unusual activity, intrusion, or a suspicious data point that lies at an irregular distance from a population. The definition of an outlier event, however, is subjective and depends on

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00020-2

the application and the domain (agriculture, healthcare, wireless network, etc.). It is important to detect outlier events as carefully as possible to avoid infrastructure failures, because anomalous events can cause minor to severe damage to infrastructure. For instance, an attack on a cyber-physical system, such as a microgrid may initiate voltage or frequency instability, thereby damaging a smart inverter, which involves very expensive repairing. Unusual activities in microgrids can be mechanical faults, behavior changes in the system, human or instrument errors or a malicious attack. Accordingly, and due to its variability, outlier detection (OD) is an ever-growing research field. In this chapter, we discuss the progress of OD methods using AI techniques. For that, the fundamental concepts of each OD model are introduced via multiple categories. Broad range of OD methods are categorized into six major categories: statistical-based, distance-based, density-based, clustering-based, learning-based, and ensemble methods. For every category, we discuss recent state-of-the-art approaches, their application areas, and performances. After that, a brief discussion regarding the advantages, disadvantages, and challenges of each technique is provided with recommendations on future research directions. This survey aims to guide the reader to better understand recent progress of OD methods for the assurance of AI.

Keywords

Outlier detection, AI assurance, ensemble learning, outlier tools, data management

Highlights

- A comprehensive review of outlier detection algorithms from the perspective of Artificial Intelligence (AI)
- Multiple outlier detection categories are introduced and relevant studies are reviewed
- Advantages, disadvantages, research gaps, and suggestions are addressed for each outlier detection category
- AI assurance is defined and discussed in relation with outlier's detection and analysis

7.1 Introduction and motivation

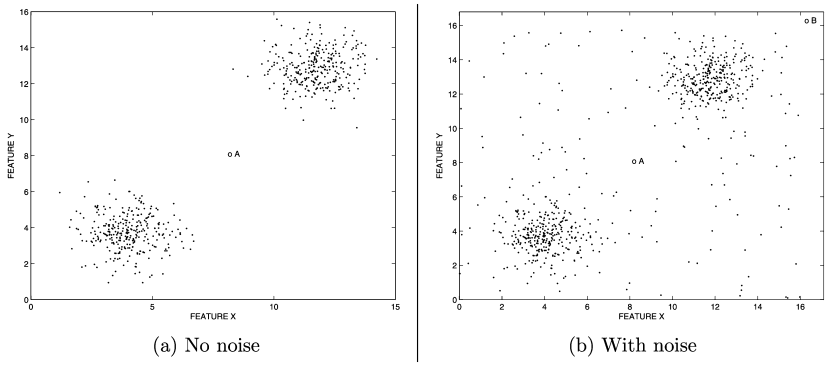
An outlier or anomaly can be defined as abnormality, deviant, or discordant data point from the remaining dataset in data science literature. According to (Hawkins, 1980, pp. 1), *"an outlier is an observation which deviates*

so much from the other observations as to arouse suspicions that it was generated by a different mechanism." During the development of AI-based applications, data are being created by several generational processes or observations collected from one or multiple entities. Outlier points generate when one or a collection of entities which behave in an unusual manner. Therefore it is very important to understand the behavior of outliers to diagnose a system's health and predict potential system failures. Some of the most popular OD applications are intrusion detection methods (Alrawashdeh and Purdy, 2016), credit card fraud detection (Porwal and Mukund, 2018), medical diagnosis (Gebremeskel et al., 2016), sensor events in critical infrastructure, precision agriculture, earth science, and law enforcement (Bordogna et al., 2007). One of the recently successful example applications of OD is credit card fraud identification, where an AI algorithm is used to find if sensitive information, such as customer identification or a card number is fraudulent or stolen. In such contexts, unusual buying patterns are observed, especially large transactions or irregular buying activities.

In networking and the Internet of things (IoT) domain, sensors are frequently used to detect environmental and geographical information; changes in underlying patterns, if they occur suddenly, might indicate important events. Event detection in sensor networks is one of the most compelling applications in cyber-physical system. Another OD example is from medical diagnosis, where data are collected from numerous medical devices, including MRI (magnetic resonance imaging) scans, PET (positron emission tomography) scans, and ECG (electrocardiogram) time-series, where unusual patterns could indicate an illness.

In data mining literature, normal data are also known as "*inliners*" (Aggarwal, 2017). Often in real-world applications, such as fraud or intrusion detection system, outliers are *sequential* and not single datapoints within a sequence. For instance, network intrusion is an event in a sequence that is intentionally caused by an individual. Properly identifying the anomalous event helps to handle those sequences.

In most conventional cases, OD algorithms have two types of outcomes: binary labels and outlier scores. Outlier scores impose the level or degree



(a) No noise (b) With noise

FIGURE 7.1 Anomalies and noise in data. Image source: (Aggarwal, 2016).

of "*outlierness*" of each data point. Scores naturally rank outlier points and provide various information about the algorithm. However, they don't represent a concise summary with small group sizes. Binary labeling is used to represent if a datapoint is a strong outlier or an inliner. Algorithms can directly provide binary labeling or other means of labeling, such as outlier scores, which then can be converted to binary labels for learning purposes. For that, a threshold is selected based on the statistical distribution of the dataset. Binary labels provide less information regarding the degree of outlierness, however in most applications, it is the desired outcomes for decision-making process.

For an outlier, defining how much *deviation* is sufficient from a normal datapoint is a subjective judgment. Datasets from real applications might contain embedded noise, and analysts might not be interested in keeping such noise. Therefore investigating significant deviation is a prime decision to make for OD algorithms. To comprehend this problem clearly, Fig. 7.1(a) and 7.1(b) illustrate two-dimensional feature spaces. It is evident that clusters are identical in both figures. However, considering a single datapoint "A" in Fig. 7.1(a) seems different from the rest of the datapoints. Therefore "A" in Fig. 7.1(a) is clearly an outlier. However, point "A" in Fig. 7.1(b) is surrounded by noise and it's quite difficult to say if it is noise or an outlier. When designing algorithms, normal and outlier boundary conditions need to be precise and specific to application requirements.

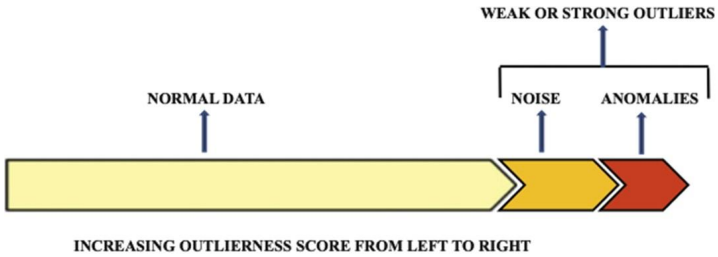


FIGURE 7.2 A typical data spectrum with noise and outliers.

In unsupervised learning models, noise is defined as weak anomalies that don't hold criteria of being an outlier. For instance, datapoints close to the boundary are mostly considered noise (as presented in Fig. 7.2). Often the separation criteria of these datapoints is subjective and depends on the interest of application-specific demands. Real datapoints that are generated from noisy environments are difficult to detect using scores. That is because noise represents deviated datapoints, therefore requires domain experts to select the threshold between noise and outliers to satisfying application requirements.

Success in OD depends on data modeling, where every application has its own unique data management requirements. Evidently, the OD technique needs to process the attribution in the data and be sensitive enough to understand the underlying data distribution model. By properly examining the data model, contextual outliers can be achieved. Aggarwal et al. (2011) proposed a concept of linkage outlier by analyzing social networks. Here, nodes that don't show any connection with each other are likely to be outliers, therefore data distribution models play an important role for designing OD models.

OD is a creative process; many researchers are trying to answer the question of how to identify outliers. Research communities are trying to bring forward many innovative and novel algorithms for OD (Aggarwal, 2017; Hadi et al., 2009). While identifying and removing outliers from the dataset, researchers need to be very observant, because sometimes outliers carry important hidden information about data. It is crucial to understand data types applying OD methods; for instance, data can be univariate or multivariate and need different approach to begin with. In statistical analy-

sis, careful observation regarding feature selection needs to be considered, because we usually want the feature to represent the data distribution model for both non-parametric and parametric analysis. Moreover, during OD, one must make analytic arguments and intuitions before making any conclusions. Besides, real world applications require context-aware and purpose-based detection, because the outcome of the result should benefit the requirements of outlier analysis in any given domain. Some recent state-of-the-art application areas are as follows:

Fraud and intrusion detection: Intrusion detection is performed to check if a computer network has any unauthorized access by observing unusual patterns (Singh et al., 2010). Additionally, to make a network secure and safe, detection of outlier instances is extremely important.

Database and sensor network monitoring: Sensor networks require continuous monitoring for effective wireless operations. Detecting outliers in sensor network (Abid et al., 2017; Feng et al., 2017), body sensor networks (Zhang et al., 2016), and target tracking environments (Shahid et al., 2015) ensures flawless operations with proper routing in the network.

IoT and critical infrastructure operations: IoT devices utilize wireless sensors to collect various information on architecture, including smart grid, power distribution system, water supply system, and healthcare diagnostic system. It's very crucial to know correct and effective data are being collected from IoT devices. If the data are being polluted with outliers because of a sensor fault or a cyber-attack, that should be identified for securing the critical infrastructure. Additionally, OD algorithms need to be trained against attack concealment. Critical infrastructures are the backbone of society; effective and efficient OD models are crucial for optimal operations, preventive maintenance, and the overall safety and security of our nation.

Data streams monitoring: Zheng et al. (2016); Tamboli and Shukla (2016); Shukla et al. (2015); Tran et al. (2016); Gupta et al. (2014); and Cateni (2008) showed OD for data streams and time series datasets. Detecting outliers in data streams is important, because any abnormality may hinder fast computational and estimation processes of applications.

Medical diagnosis: Modern healthcare and diagnosis analysis are mostly dependent on electronic devices. These devises observe unusual patterns while reading different measures from patients. Properly separating anomalous readings help doctors to predict underlying conditions, and thereby to apply proper diagnosis.

Fake news detection: Fake news can be considered as an outlier if compared with foundational datasets and real sources (Shu et al., 2017).

Surveillance and security: Security is an important aspect in computer administrative network. Cybersecurity is a field where researchers ensure methods for safe access and proper authentication. An exciting and practical research in cybersecurity is surveillance video OD (Xiao et al., 2015).

Data logging and data quality: Logging and processing data for commercial purposes can go wrong because of unwanted concealment processes, which if not detected, might result in irrecoverable loss. Automated data mining models are applied in searching for abnormalities while processing large volume of logs (Ghanbari et al., 2014). Proper anomaly identification algorithms need to be applied to enhance data quality (D'Urso, 2016; Chenaoua et al., 2014).

The rest of the chapter is organized as follows: In Section 7.2, we categorize OD algorithms into six subgroups, where each subgroup has a detailed discussion, advantages, disadvantages, research gaps and suggestions. In Section 7.3, we include multiple OD tools. In Section 7.4, we enlist several benchmarking datasets for outlier analysis, and in Section 7.5, we discuss AI assurance and its relevance to outlier analysis. Finally, in Section 7.6, we conclude with open research gaps and OD challenges.

7.2 Outlier detection methods

OD methods can be classified into many categories (Ranshous et al., 2015; Braei and Wagner, 2020; Lai et al., 2020), however, in this chapter we introduce six major categories: Statistical, Density, Clustering, Distance, Learning, and Ensemble-based OD methods. For each group, we provide short overview about their gradual development over the last few decades.

7.2.1 Statistical and probabilistic based methods

Statistical and probabilistic-based OD methods originated from early nineteenth century (Edgeworth, 1887). Before inventing high performance devices these methods were applied for simple data visualization, although performance and efficiency were being neglected. Nevertheless, the fundamental mathematics are always useful and eventually these methods are applied to most regular OD applications.

Almost all the OD algorithms apply numerical scores to every object and in the final step they assign extreme values by observing the scores. Binary classification is one way of sorting the extreme value points. Statistical and probabilistic OD algorithms can be either supervised, unsupervised or semi-supervised. The model is built based on data distribution. For statistical-based OD algorithms, stochastic distribution is a widely adopted technique to detect outliers. Therefore the degree of outlierness depends on the model built using data distribution. Statistical and probabilistic-based methods can be further divided into two broad categories: parametric and non-parametric distribution models. Parametric methods assume a distribution model from the dataset, and then use knowledge from the data to approximate model parameters. Non-parametric methods don't assume any underlying distribution model (Eskin, 2000).

7.2.1.1 Parametric distribution models

Parametric distribution models have prior knowledge of the data distribution, these models can be divided into two subcategories: Gaussian mixer and regression models.

Gaussian mixture models: Gaussian model is a popular statistical approach in OD, it initially adopts maximum likelihood estimation (MLE) in training stage to compute variance and mean of the Gaussian distribution. During the test phase, several statistical measures are applied (mean variance test, box plot test) to validate the outcomes.

Yang et al. (2009b) proposed an unsupervised Gaussian mixture model (GMM) based on an explainer that globally optimizes to detect outliers. In this method, first it fit the GMM for a dataset by utilizing the expectation maximization (EM) algorithm based on global optima. Outlier factor

for this method is calculated as the sum of proportional weighted mixture, the weights represent affiliations to remaining datapoints. Mathematically, outlier factor can be expressed as x_k:

$$F_k = z_k(t_h) = \sum_{j=1}^{n} s_{kj}\pi_j(t_h) \tag{7.1}$$

where, $s_{kj}\pi_j(t_h)$ = Point X_k's relationship with other point X_j.

s_{kj} = Relationship strength

t_h = Iteration (Final)

π_j = Degree of importance of point j

Higher outlier factor indicates greater degree of outlierness. This method focuses on global properties rather than local ones that we discuss later in density-based method section (Breunig et al., 2000; Papadimitriou et al., 2003; Tang et al., 2002). Yang et al. (2009b) claimed, for a given dataset, fitting the GMM at each data point outlier can be detected, even if the dataset contains noise, which was a major challenge in clustering-based techniques. Therefore this technique is useful in real-world applications, where environmental noise or intentional adversarial noise is embedded. It is evident that the algorithm has higher capacity to detect unusual objects, however, it incurs greater complexity: for single iteration model complexity is $O(n^3)$ and for N iteration model complexity is $O(Nn^3)$. Future studies shall improve the algorithm and reduce its computational complexity along with increasing its scalability.

Tang et al. (2015) proposed an improved and robust statistical model; they applied GMM with projections preserving locally. They applied the model to disaggregate energy utilization by combining both outcome of subspace learning (SL) and GMM. In this method, the LPP short for locality preserving projection of SL is exploited to reveal the inherent diverse structure, while at the same time keeping the neighborhood composition intact. Saha et al. (2009) proposed a principality component analysis (PCA) technique that points research gaps in local outlier factor (Breunig et al., 2000) and connective-based outlier factor (Tang et al., 2002) that fails to achieve multi-Gaussian and multiple state OD. The method shows im-

proved performance, however, the authors barely discussed anything about their model's computational complexity.

Regression models: Regression OD models, depending on the context, are either linear or non-linear. They are a direct approach to detect outliers. Generally, the training stage involves fitting the given datapoints into a constructed regression model. The regression models are evaluated at the test stage. Outliers are labeled if the difference between actual output and predicted outcome of the regression model is too high. For the last few years, OD using regression analysis applied several standard techniques as Mahalanobis distance, mixture models, robust least squares, and Bayesian alternate vibrational methods (Zhang, 2013). Satman (2013) in contrast to other algorithms, proposed a different one, one that has a covariant matrix that is non-interactive. It has less computational complexity, which makes it cost effective as it can detect multiple outliers quickly. For future research directions, and as regression models often portrayed as minuet preference, variance and bias of the intercept approximator can be minimized to improve the result.

Another regression model proposed by Park and Jeon (2015) detects outliers in sensor network. The method observes the values from the model outcome and create an independent variable using a weighted sum approach. Since the model only applied on a single sensor environment, measuring outliers accurately from multiple sensor environment can be an interesting topic (as a future direction). Dalatu et al. (2017) studied a comparison between linear and non-linear model, where their accuracy and misclassification were examined with receiver operating characteristic (ROC) curves. This case study provided necessary information for OD for two popular kinds of regression models. Non-linear models showed more accuracy (accuracy 93%) compared to linear regression models (accuracy 63%), therefore it's mostly a better option to select non-linear models over linear regression models.

7.2.1.2 Non-parametric distribution models

Non-parametric distribution models don't assume any underlying data distribution (Eskin, 2000) for given datasets. Kernel density estimation (KDE)

models are a popular non-parametric approach; they are unsupervised technique to detect outliers that utilizes kernel functions (Latecki et al., 2007). The KDE model compares each objects' density with neighbors' densities, where the idea is similar as some of the prevalent density-based techniques (Papadimitriou et al., 2003; Breunig et al., 2000). Although, it has improved performance, the curse of dimensionality reduces its applicability. Gao et al. (2011) offered a superior solution to overcome the problem. They applied kernel-based technique that has lower run time compared to (Latecki et al., 2007; Breunig et al., 2000), also presented better scalability and performance for large datasets. This method solves another limitation of the local outlier factor (Breunig et al., 2000): sensitivity on parameter k, where it measures the weights of local neighborhoods by utilizing weighted neighborhood density estimations.

A good real-world application by Samparthi and Verma (2010) also applied KDE to measure infected nodes in a sensor network. Boedihardjo et al. (2013), in another study, implement the KDE method in time series dataset, although it was a challenge using KDE for data streams. They proposed an accurate estimation of probability density function (PDF) by using adaptive KDE. The computational cost associated with the method is $O(n^2)$, and showed better estimation results compared to original KDE. The method is suitable for strict environment, therefore further research may improve the method for adopting multivariate data. Uddin et al. (2015) applied the KDE method in power grid environment. Although, the KDE methods are better at targeting outliers, they are computationally expensive. Later, Zheng et al. (2016) applied KDE in a multimedia network for outlier detection on multivariate dataset. In another study, Smrithy et al. (2016) introduced a non-parametric method for outlier detection in big data. Later, an adaptive kernel density-based approach, a nonlinear method, based on Gaussian Kernel, is proposed by Zhang et al. (2018). Later, Qin et al. (2019) proposed a unique OD approach that perfectly applies KDE to effectively identify local outliers from continuous datasets. This method facilitates to detect outliers from high data stream, irrespective of data complexity and unpredictable data update, which was a challenge earlier. They derived an approach to successfully identify top-N outliers based on KDE on continuous data. Af-

terwards, Ting et al. (2020) modified the KDE approach to identify similarity between two distribution named isolation distribution kernel. Compared to other kernel-based algorithm, the proposed method outperforms most point anomaly detection. Although, KDE-based approach performs better compared to other non-parametric models, they suffer from high dimensionality in the feature space. Additionally, in general, they have high computational cost too.

7.2.1.3 Miscellaneous statistical models

Among many proposed OD algorithms, most straightforward techniques in statistical method are trimmed mean, boxplot, Dixon test, histogram, and extreme studentized deviate (ESD) test (Goldstein and Dengel, 2012; Walfish, 2006). The Dixon test works well with small size dataset, as no assumption is required about data normalcy. The trimmed mean is not a good approach among all others for OD, however, ESD test is a better choice. Pincus (1995) introduced several optimization tests for OD that could depend on parameters such as number and expected space of outliers. A histogram-based OD technique, HBOS (histogram-based outlier) is proposed by Goldstein and Dengel (2012), which can create model of univariate feature space by utilizing dynamic and static histogram bin width. Here, each data point is scored as degree of outlierness. The algorithm showed improved performance, especially faster computational speed over traditional OD approaches (Jin et al., 2006; Tang et al., 2002; Breunig et al., 2000). Nevertheless, the method faces difficulties finding local outliers with its density approximation technique.

Hido et al. (2011) introduced a novel statistical methodology by applying guided density ratio approximation to detect outliers. The main idea of the algorithm is to select density ratio between training set and test set. A natural cross validation method was applied to optimize the value of parameters: regularization and kernel width. To achieve better cross validation performance, unconstrained least square method was applied. This method has an advantage over non-parametric kernel density estimation, because hard density estimation isn't required here. The method, in terms of accuracy, shows improved performance in most cases. Improving density ratio estimation of this method is an important research direction.

Robust local outlier detection (RLOD), another method that adopts statistical measures to detect outliers is proposed by Du et al. (2015). This pipeline assumes the fact that OD is sensitive to parameter tunning (Gebhardt et al., 2013) and most OD methods are focused to detect global outliers. The whole pipeline can be divided into three stages. At the first stage, it applies three standard deviation measures to find density peaks of the dataset. In the 2nd stage, remaining data points are labeled to the closest higher density neighbors by assigning them in matching clusters. In the 3rd and final stage, it applies density reachability and Chebyshev's inequality to locate local outliers for each collection. Campello et al. (2015) showed that RLOD can both detect local and global outliers; they experimentally showed that RLOD outperforms some former OD algorithms (Breunig et al., 2000; Zhang, 2013) in terms of detection rate and execution time. RLOD performance can be improved more by adopting parallel and distributed computing. Later in another study, Li et al. (2020) proposed an effective copula-based OD.

7.2.1.4 Advantages of statistical and probabilistic based methods

The fundamental mathematics behind statistical OD algorithms make them easy to use. Due to their compact form, the models exhibit improved performance in terms of detection rates and run times for a particular probabilistic technique. For quantitative ordinal and real-valued data distribution, the models usually fit well, although results could be more improved if ordinal data can be preprocessed. Despite some targeted issues, such as high dimensional feature space, the models are convenient to deploy.

7.2.1.5 Disadvantages of statistical and probabilistic based methods

The parametric models assume underlying density distribution, which results in poor performance and often might bring unreliable outcomes in real-world applications, such as managing data streams from a complex network. Statistical-based approach is applicable mostly for univariate datasets; therefore they don't perform well for multivariate feature spaces. If the models are applied to multivariate feature space, high computational cost incurs, which make them a poor choice for multivariate data stream. Additionally, the histogram cannot capture the interaction between fea-

tures, which makes it a poor choice for high-dimensional data as well. Therefore statistical methods that can investigate simultaneous feature space can be promising research. To deal with the curse of high dimensionality, specific statistical methods can be adopted, however, it results in longer processing time and a misleading data distribution.

7.2.1.6 Research gaps and suggestions

Several common research gaps in statistical-based approach are poor accuracy, difficulties with high-dimensional datasets, and operational expense. These gaps need to be addressed in the future to make the models more reliable. These methods, however, can be more effective if the applied model is aware of the context. Time series data generated from critical infrastructures, such as smart grid and water distribution system, may contain anomalous samples because of maintenance problems or intentional attacks, however, their pattern is unknown to a learning model. In this scenario, parametric methods fail to learn the underlying distribution, as it constructs the model based on predefined data distribution. Therefore for this case non-parametric methods are a better choice, as they don't need to know the underlying distribution of a given dataset. Also, parametric methods are not a better choice for large data stream, where outlier points are dispersed evenly. Inaccurate labeling of outliers might occur if the threshold is defined based on standard deviation to separate them. Using parametric methods for OD is a difficult task while applying GMM to manage data stream and high-dimensional feature space. Therefore algorithms that can easily manage data stream along with high-dimensional feature space can make the model more scalable. High dimensionality also creates problem for regression models. To overcome this issue, targeted regression analysis can be adopted instead of ordinary regression analysis.

Non-parametric models, especially KDE are a better choice in most applications, however, they get computationally expensive in noisy environments. In contrast with parametric methods, KDE is scalable, although computationally expensive for multivariate data. The histogram-based approach is a good fit for univariate data distribution, however, its inability to investigate the interaction among features makes it a poor choice for

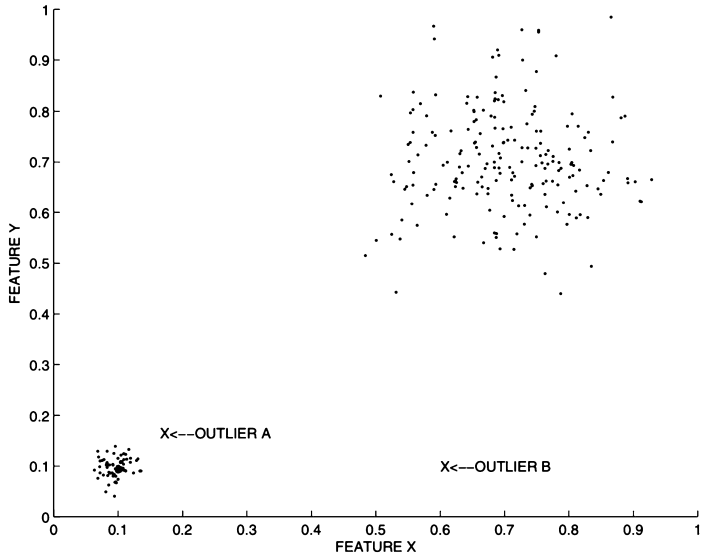


FIGURE 7.3 Density-based outlier detection (Aggarwal, 2016).

multivariate data. Despite statistical methods inability to adopt some recent application areas, they are still a good choice for targeted domain and data streams. PCA methods by Saha et al. (2009) and Tang et al. (2015) are effective approaches for OD. Goldstein and Dengel (2012) proposed a histogram-based outlier (HBOS); it shows improved performance when compared to other clustering-based models, such as local outlier factor, local correlation integral, and influenced outlier in terms of calculation speed, therefore is a good choice for real-time data (Breunig et al., 2000; Papadimitriou et al., 2003; Jin et al., 2006). OD models scalable to large dataset proposed by Du et al. (2015) and Hido et al. (2011) also proved robust in analyzing outliers.

7.2.2 Density-based methods

Density-based OD is one of the most popular and prevalent techniques. The main principal is that an outlier point can be found in a sparse region, whereas normal points can be found in denser region. Fig. 7.3 presents a two-dimensional dataset, where labeled point "A" and "B" are considerably separated from the rest of the densely populated clusters, therefore are

outlier points in this dataset. The core idea for detecting outlier points "A" and "B" is that these points remain in sparse populations, whereas the normal points are in higher denser populations. Density-based methods seek for differences between densities of a point with their local neighborhood. Usually, density-based methods are computationally expensive compared to distance-based methods. Despite this problem, density-based methods are widely popular because of their simplicity and efficiency to detect outliers. Some baseline algorithms utilizing these methods are presented in Breunig et al. (2000); Jin et al. (2006). Zhang et al. (2009b); Tang and He (2017) presented algorithms that are developed and modified version of those baseline one's.

Local outlier factor (LOF): LOF is a popular method proposed by Breunig et al. (2000), which is the base algorithm that represents density-based clustering method for detecting outliers. K-nearest neighbor (KNN) technique is used in this process for each point in a KNN set. LOF measures local reachability density (lrd) to differentiate each point with its neighborhood. Mathematically, lrd can be defined as

$$lrd\,(p) = \frac{1}{\frac{\sum_{o \in kNN(p)} reach-dist_k(p \leftarrow o)}{|kNN(p)|}} \tag{7.2}$$

$$\text{LOF score: } LOF_k\,(p) = \frac{1}{|kNN\,(p)|} \sum_{o \in kNN(p)} \frac{lrd_k\,(o)}{lrd_k\,(p)} \tag{7.3}$$

where, $lrd_k\,(p) = lrd$ of point p

$lrd_k\,(o) = lrd$ of point o

The main idea of the LOF is to determine the degree of outlierness of an observation, while comparing its cluster with local neighbors. The LOF score gets higher for an observation if its lrd value is less than the estimated nearest neighbor. Logging lrd value and computing LOF score using KNN approach costs $O\,(k)$ for each data point. It is wise to use a valid index, because a sequential search of a size n dataset can cost n^2 if a proper indexing is not applied.

Schubert et al. (2014) addressed this shortcoming and introduced a simplifiedLOF method, which makes the density estimation simpler. The sim-

plifiedLOF method adopts KNN distance instead of LOF's reachability distance.

$$dens\,(p) = \frac{1}{k - dist\,(p)} \tag{7.4}$$

The simplifiedLOF is more computationally complex than LOF, but improved in performance.

Connective-based outlier factor (COF): Tang et al. (2002) realized an improved method, COF, over methods proposed by Breunig et al. (2000); Schubert et al. (2014). The COF is almost similar to the LOF, although density estimation calculation is different. The COF applies chain distance to calculate local densities of neighbors, but Euclidean distance is generally applied to LOF. Because of applying chaining distance for density estimation, this process assumes predefined population distribution, which is a major drawback, because it often results in wrong density estimation. The authors applied a new term- "isolativity" instead of "low-density" to locate outliers. Isolativity is a unique measure that represents the degree of connectedness of an observation with the remaining points. At point p, the COF value can be expressed mathematically, while applying the KNN approach is

$$\mathrm{COF}_k(p) = \frac{\left|N_{k(p)}\right| ac - dist_{Nk(p)}\,(p)}{\sum_{o \in Nk(p)} ac - dist_{Nk(p)}\,(p)} \tag{7.5}$$

where $ac - dist_{Nk(p)}\,(p) =$ Average chain distance between point p and $N_{k(p)}$.

In the neighborhood, COF modifies density estimation of the SimplifiedLOF to verify the connectedness using a method called minimum spanning tree (MST). The computational cost is $O(k^2)$ that occurs for calculating MST from KNN set. Except in circumstances, where datasets are characterized by connective data patterns, COF takes similar time as LOF for detecting outliers.

Local outlier probabilities (LoOP): The LOF algorithm uses scores for each datapoint of KNN. However, threshold selection for labeling datapoints was a growing question. Therefore Kriegel et al. (2009b) proposed LoOP that generates score with statistical probabilistic approach. In this method, density is estimated using distance distribution. LOF scores are presented as

statistical probabilities. They compare the advantages of assigning probabilities of a datapoint over outlier score in LOF. Mathematically LoOP can be expressed as

$$LoOP_s(O) = max\left\{0, erf\left(\frac{PLOF_{\lambda,S}(O)}{nPLOF.\sqrt{2}}\right)\right\} \tag{7.6}$$

where, $PLOF_{\lambda,S}(O) =$ LOF probability wrt importance of λ, r

$nPLOF =$ aggregated value

Normal points that are in denser population will have LoOP value almost zero, whereas LoOP value towards 1 indicated loosely connected points or outliers in the dataset. Just as simplifiedLOF (Schubert et al., 2014), the LoOP also has same computational complexity for each point: $O(k)$. A significant difference for calculating local densities compared to previous density-based methods is that it assumes and applies half-Gaussian distribution for density estimations.

Local correlation integral called (LOCI): Papadimitriou et al. (2003) proposed a method called LOCI that correctly handles multi-granularity issue, where LOF (Breunig et al., 2000) and COF (Tang et al., 2002) both were unable to solve the problem. They defined an outlier metric-MDEF, short for multi granularity deviation factor. According to the method, outliers are points that are away from the neighbor of MDEF by at least three standard deviations. Not only does this method find both remote cluster and concealed outliers, but also deals with feature space local density variation. The MDEF can be defined mathematically on a point p_i within a radius r:

$$MDEF(p_i, r, \alpha) = 1 - \frac{n(p_i, \alpha r)}{\hat{n}(p_i, r, \alpha)} \tag{7.7}$$

where, $n(p_i, \alpha r) = \alpha r$ neighborhood objects number

$\hat{n}(p_i, r, \alpha) =$ All the objects p's average at r-neighborhood of p_i

To get faster result from MDEF, the right side fraction needs to be measured after getting the value of numerator and denominator. So far, all the OD algorithm we have discussed are based on KNN algorithm; detection

of number of k is crucial to find outliers properly. The LOCI algorithm is better, because it applies a maximization process to find out optimal k-value. It is because for estimating local densities LOCI applies half Gaussian distribution that mimics LoOP (Kriegel et al., 2009b). Instead of measuring distances for density estimation, they aggregate the local neighborhood records. Also, LoOP is different, because it of its unique neighbor comparison. Although LOCI shows good results, it has longer run time. Papadimitriou et al. (2003) developed a different approach to increase the speed by introducing quad tree with several constraints between two neighbors.

Relative density factor (RDF): Ren et al. (2004b) proposed a new technique for effective OD by pruning datapoints located in deep cluster. This algorithm takes advantage of large datasets and provides scalability. RDF adopts a data model to identify anomalies, called *P-tree*. Higher RDF values indicate greater outlier behavior of datapoints in the population. RDF can be mathematically expressed on point p and radius r as

$$RDF(p,r) = \frac{DF_{nbr}(P,r)}{DF(P,r)} \tag{7.8}$$

where, $DF_{nbr}(P,r)$ and $DF(P,r)$ are both density factor

Influenced outlier (INFLO): INFLO is another technique based on LOF (Breunig et al., 2000) and proposed by Jin et al. (2006). The method detects outliers by assuming symmetric relationship between neighbors. One shortcoming of LOF (Breunig et al., 2000) is that it fails to correctly define scores for datapoints at cluster border, where the clusters are related closely. INFLO solves this problem by distinguishing different neighborhood of context and reference set. The scores are calculated by both reverse nearest neighbor and KNN. INFLO adopts both reverse nearest neighbors and nearest neighbors techniques to achieve accurate neighborhood distribution. Here outliers are observations that have higher INFLO values.

High contrast subspace (HiCS): Almost all the previous algorithms described (LOF, COF, LOCI, and INFLO) suffer when calculating distances of large dimensional feature spaces. However, a method proposed by Keller et

al. (2012) for large dimensional dataset can successfully sort and rank outliers and their score: high contrast subspace method (HiCS).

Global-local outlier score from hierarchies (GLOSH): Campello et al. (2015) proposed a method that includes beyond local outliers and extends the investigation to detect global outliers. This method applies statistical interpretation to find both local and global outliers. Although GLOSH isn't a generic algorithm, often it provides better results. The baseline algorithm is KNN; therefore it has some common setbacks, which can be solved by further improving density estimation.

Dynamic-window outlier factor (DWOF): Momtaz et al. (2013) proposed a unique algorithm that detects top n number of outliers by assigning outlier score called DWOF. This method deviates from its ancestor algorithms in density-based methods. However, it closely complements Fan et al. (2009) proposed resolution-based outlier factor (ROF). ROF performs better in terms of accuracy and sensitivity to hyperparameters.

Algorithms for high-dimensional data: With the increment in data volume and complex networks, its highly required to design sophisticated and efficient algorithms. Keeping that in mind, Wu et al. (2014) implemented an algorithm that can handle high-dimensional data. The algorithm introduces a new technique, called RS-forest, which is faster and more accurate. It includes one class semi-supervised machine learning (ML) model. Later, Bai et al. (2016) proposed a similar technique as Wu et al. (2014), which can discover outliers in parallel. LOF (Breunig et al., 2000) is the base algorithm, but a new computing method is introduced, called distributed computing for density estimation. This algorithm works in two steps: at first it partitions using grid-based technique, and then distributed computing identifies the outliers. Unfortunately, this algorithm doesn't scale well; earlier Lozano and Acuna (2005) fixed this issue by suggesting a technique called PLOFA (parallel LOF algorithm), which improves scalability for big data.

Other density-based algorithms: Tang and He (2017) proposed a method to estimate density using kernel density estimation for measuring local anomalies; a scoring process is introduced, called relative density-based

outlier score. The model applies KDE that pays more attention on shared neighbor and reverse neighbors, rather than KNN to compute density distribution. Distance measure is the same as UDLO (Cao et al., 2014a), which is Euclidean distance. However, they need to compare different distance measures to observe the changes before applying this method to real applications. Iglesias Vázquez et al. (2018) introduced a detection algorithm for data that have low density population, called sparse data observation (SDO). The SDO is a hungry learning algorithm that tries to learn quickly and reduces computation time for each object compared to previous algorithm in density-based methods that we have discussed so far. Ning et al. (2018) proposed relative density-based OD method, which is a similar method to Tang and He (2017); it's a new technique to compute neighborhood density distribution. Su et al. (2019) implemented local OD algorithm on scattered dataset, instead of using the term LOF, they used local deviation coefficient (LDC), because the LDC focuses on distribution of object and neighbors. The algorithm removes normal points in a safe way and keeps the outlier points as reminder; the process is called RCMLQ (rough clustering based on multi-level queries. Since, it prunes the normal objects, it is useful for local OD in large dataset. It showed better efficiency and accuracy over previous local OD algorithms.

7.2.2.1 Advantages of density-based methods

Density-based OD algorithms apply non-parametric method to measure density, therefore they don't assume any predefined distribution model to manage the dataset. LOF (Breunig et al., 2000), LoOP, INFLO (Jin et al., 2006), and DWOF (Papadimitriou et al., 2003) are some of the baseline algorithms that serve as the fundamental model. Density-based algorithms can both identify local and global outliers, which make them useful for real-world application and often outperform other statistical-based algorithms (Wang et al., 1997; Akoglu et al., 2014; Hido et al., 2011). Additionally, the fundamental concept is to estimate neighborhood density that provides more flexibility to investigate crucial outliers, which can be easily measured by several other modern OD algorithms. Density-based algorithms also facilitate excluding outliers from nearby denser neighbors. They hardly require any primary knowledge, such as probability distribution, which

makes the algorithm easy for hyperparameter tuning. In fact, only single hyperparameter tuning brings good results. The algorithms are also useful and efficient when it comes to detecting local outliers (Su et al., 2019).

7.2.2.2 Disadvantages of density-based methods

Although some of the density-based algorithm showed good performance, they are computationally expensive and complicated when compared to many statistical-based methods, including ones presented by Kriegel et al. (2009a). Also, these methods are sensitive to the shape of the neighbors; when cautiously tuning the size hyperparameter, they become computationally expensive, including increased runtime. It is also evident from the applications that neighbors varying density creates complicated models and generally generates poor result. Few density-based methods, such as MDEF and INFLO, because of their complex density estimation process, cannot handle datasets resourcefully, such as defining outleirness of an object. Also, density-based models face challenge when it comes to managing high-dimensional time series data. However recent algorithms seem to overcome the problems by introducing pruning (Ren et al., 2004b) and elimination (Su et al., 2019) techniques, among others.

7.2.2.3 Research gaps and suggestions

In general, since density-based OD's are non-parametric methods, sample size is considered small for high-dimensional feature space. This challenge can be resolved by resampling the objects to enhance the process. As density-based algorithms are based on k-nearest neighbors, therefore proper selection of hyperparameter k is important to evaluate these algorithms. Generally, computational expense using KNN is $O(n^2)$. However, LOCI has greater complexity because of adding an extension, radius r; therefore computational cost becomes $O(n^3)$. So, LOCI, when applied to big data, gets very sluggish to compute OD. Goldstein and Uchida (2016) compared LOF and COF. They concluded that applying spherical density estimation using LOF creates a poor-quality process for OD. However, COF applies connectivity feature to estimate density pattern to solve the issue. INFLO, when applied to closely related clusters with varying densities, performs better by generating enhanced outlier scores.

7.2.3 Clustering-based methods

Clustering-based OD differentiates between clusters and outlier points. A simple description would be: *each datapoint in a given dataset that belongs to a cluster is either an outlier or a normal point*. The goal for clustering is to separate the points from denser and sparse population; generally, a sparse region contains most of the outliers. Therefore most clustering algorithms get outliers as a side product of their analysis. While detecting outliers using clustering-based approach, a score is provided that represents the degree of outlierness of a sample. Outlier score can be calculated using the distance between a datapoint and nearest cluster centroid. Because of different cluster shape, Mahalanobis is a good distance measure that scale well for the clusters. Mathematically, Mahalanobis distance from datapoint X to cluster distribution with centroid μ and covariance matrix Σ is

$$MB(X, \mu, \Sigma)^2 = (X - \mu)\, \Sigma^{-1} (X - \mu)^T \tag{7.9}$$

Here, $X = dataset$,

$\Sigma = covariance\ matrix$,

$\mu = attribute\ wise\ means\ of\ d\ dimensional\ row\ vector$

After scoring each datapoint with the Mahalanobis distance, binary labels can be assigned by selecting extreme comparison. Mahalanobis distance can be visualized as the Euclidean distance between a sample and a cluster centroid. This distance measure indicates data locality characteristics by providing statistical normalization.

Fig. 7.4 illustrates the effects of identifying outliers, while considering data locality. Here, Euclidean distance measure will consider point "A," an outlier over point "B," because of the normal distance measure. However, Mahalanobis distance, considering data locality, provides point "B" as more anomalous than point "A," which makes sense visually (Fig. 7.4). Therefore defining a proper number of clusters and a suitable distance measure results in successful outcome of the OD algorithm.

Detecting outlier using clustering-based approach is dependent on properly defining cluster structure of normal instance (Al-Zoubi, 2009), which comes from the effectiveness of the algorithm. These algorithms

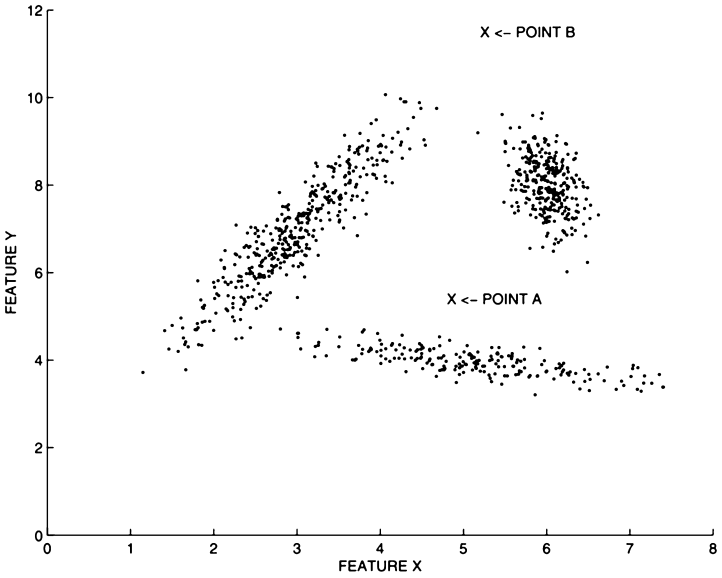


FIGURE 7.4 Clustering-based OD method (Aggarwal, 2016).

are *unsupervised* since they don't need any previous knowledge of feature distribution. Many OD techniques are introduces based on clustering algorithms; Zhang (2013) categorized several of them. Clustering-based approach is a broad category and can be grouped into several subgroups as following:

Clustering methods based on partitioning: These clustering methods are based on distance-based technique, where cluster numbers are selected initially or provided randomly. Algorithms belong to this subgroup are presented by MacQueen (1967); Ng and Han (1994); Kaufman and Rousseeuw (2009).

Clustering methods based on density: In contrast to partitioning-based clustering approach, defining initial number of clusters for these models isn't required. However, they can model the cluster into denser and non-denser groups given the radius of a cluster. Algorithms belonging to this subgroup are studied by Hinneburg and Keim (1998), including density-based spatial clustering of applications with noise (DBSCAN) by Ester et al. (1996).

Clustering methods based on hierarchy: In this subgroup, the algorithms partition the cluster into different levels structured like a tree. Algorithms belonging to this subgroup are presented by Karypis et al. (1999); Guha et al. (2001); Zahn (1971).

Clustering methods based on grids: Algorithms belonging to this subgroup are presented by Zhang et al. (2005); Sheikholeslami et al. (2000); Wang et al. (1997).

Clustering methods based on high dimensional features: Algorithms belonging to this subgroup are presented by Agrawal et al. (1998); Aggarwal et al. (2004). Besides that, Cao et al. (2006) proposed a two-stage algorithm called DenStream. They applied density-based approach for both offline and online OD. The first stage summarizes the given time series dataset, then the second phase organizes clusters from the summarized data. The DenStream creates a microculture to separate outliers and normal data points. A micro cluster is a real outlier if its weight is less than the predefined threshold and being pruned by the model afterwards. The authors performed a comparison between DenStream and CluStream (Aggarwal et al., 2003) to present their models' effectiveness. DenStream shows improved performance, because it avoids using memory space and utilizes taking snapshots on a disk. However, the model faces difficulties when adjusting dynamic parameters in time series datasets and locating arbitrary cluster shapes with multiple levels of granularity. Solving these issues can be a good future study. Later, Chen and Tu (2007) proposed an algorithm like the DenStream, regarding offline and online OD, called D-Stream; the only difference is that D-Stream is a grid-based OD algorithm. Outliers, compared to previous algorithm, can be found easily by exploiting the definition of noise in terms of dense, sparse, and sporadic grid. A density threshold is selected to which the sporadic grids are compared, if less than the threshold the datapoints are considered outliers. Also, the algorithm performs better in terms of clustering and runtime compared to CluStream. In another study, Assent et al. (2012) implemented an algorithm called AnyOut for computing outliers from data stream anytime. The AnyOut algorithm builds a precise tree topology, ClusTree, to identify outliers at any time, whether the data are

constant or varying. ClusTree is a special feature of the model; it plays a part in creating the clusters.

A clustering-based approach using k-means was proposed by Elahi et al. (2008); it detects outliers by splitting data streams into chunks. Although the model doesn't perform well for grouped outliers. They experimentally presented following: comparison with some existing approach (Angiulli and Fassetti, 2007b; Pokrajac et al., 2007) demonstrates that the model has improved performance for investigating outliers from data streams. The authors suggested that combining distance-based methods with their clustering model will yield better results. However, the model merely discovers the outliers, but doesn't assign any outlier scores. MacQueen (1967) presented a pipeline to investigate outliers in varying data streams by utilizing similar approach as k-means. The model assigns weights for each feature based on their significance. The weighted features are significant, during algorithm processing they restrain noise effect. Comparing the algorithm with LOF (Breunig et al., 2000), it showed better detection rate, including low time dissipation and low false positive rates. However, the algorithm doesn't define the degree of outlierness; therefore it might be a good future study to extend the pipeline and make it scalable over different data types. Later in another study, Morady et al. (2013) tried to implement cluster-based algorithm for big data, applying k-means algorithm to build an advanced pipeline; it was deemed successful.

Bhosale (2014) combined both partitioning and distance-based approach to build an unsupervised model for data streams. They used partitioning clustering scheme (Ng and Han, 1994), which provides weights to the clusters according to their adaptivity and relevance by utilizing weighted k-means clustering. The concept of the model can evolve and adapt incrementally. The authors mentioned that it has higher OD rates than Elahi et al. (2008), and they suggested to include both categorical and mixed data as part of a future study. Another interesting method proposed by Moshtaghi et al. (2014) showed a clustering algorithm that can identify outlier beyond the cluster boundary. To observe the primary change is data stream distribution, the model continuously updates mean and covariance matrices. In another study by Moshtaghi et al. (2015), they proposed an-

other framework on top of their previous one (Moshtaghi et al., 2014). The authors applied elliptical fuzzy logic to model the streaming data, to identify outlier; fuzzy parameters are updated by same style as in Moshtaghi et al. (2014). For evolving dataset, Salehi et al. (2014) implemented an architecture based on ensemble learning. Ensemble methods create several clustering models instead of modeling the data streams and updating it from time to time. Evaluating all the clustering models, few are selected to measure the degree of outlierness for each datapoint. An efficient algorithm, based on clustering technique, was proposed by Chenaghlou et al. (2017). It showed improved memory usages and lower run time by presenting the concept of an active cluster. For any given data, they are divided into chunks, where active clusters are analyzed in each chuck of data; the underlying data distribution also gets revised. Rizk et al. (2015) implemented an algorithm that investigates outliers in both small and large clusters. In another study, Chenaghlou et al. (2017) modified the method to perform detection in real time by Chenaghlou et al. (2018). Additionally, the model can detect cluster evolution sequentially. An effective algorithm, a cluster text OD algorithm, is proposed by Yin and Wang (2016). If the chance of recognizing a cluster is low, it's highly probable to be an outlier. The model presents a technique (GSDPMM: Gibbs sampling of Dirichlet process multinomial mixture) to find if a document that held in a cluster is an outlier. Relating GSDPMM with incremental clustering can be a worthy research direction, as GSDPMM has a potential in incremental clustering. Later, Sehwag et al. (2021) proposed a unique framework, called self supervised detection (SSD), based on unlabeled distributions. They experimentally showed that their method, when it comes to unlabeled data, outperforms some of the traditional OD algorithms, and even performs better than supervised detectors.

7.2.3.1 Advantages of clustering-based methods

Clustering-based methods are unsupervised, therefore if underlying distribution knowledge is not necessary, then these models are a suitable choice. After the models learn about the clusters, they can test additional datapoints for detecting outliers. Again, the unsupervised nature is suitable for incremental model as underlying distributions aren't required. They are

robust algorithms and can manage versatile data types. For example, the hierarchical clustering methods for OD are good choices for different data types; they produce nested multiple partitions, which is helpful for users to select partitions belonging to a certain level.

7.2.3.2 Disadvantages of clustering based methods

A major drawback of clustering-based algorithm is that the outliers aren't assigned a score, but binary labeling, where score represents degree of outlierness for a sample. Scoring is necessary, because it helps to back track model actions; therefore the actions of a model become final and cannot be undone. Declaring the best number of clusters initially is a difficult job, and most of the clustering algorithm often face difficulties with it. Also, if the cluster shape is arbitrary, the algorithms face problems understanding exact clusters from a given dataset. Therefore to perform well, the shapes of several clusters need to be defined initially, although it is a daunting task to provide the shape and distribution of multiple clusters. Partitioning-based methods are very sensitive to initialization of parameters, such as density-based methods. Nevertheless, they are inadequate to describe clusters and in most cases are not suitable for very large dimensional datasets. Additionally hierarchical-based clustering methods showed expensive simulations in methods proposed by Karypis et al. (1999) and Zahn (1971), which makes them a poor choice for large datasets.

7.2.3.3 Research gaps and suggestions

It is important to note that, when designing any cluster-based models several questions need to be answered. In relation to an object defined as outlier: does it belong to a cluster, or is it located outside of the cluster boundary? If the distance between the object and the cluster centroid is distance, can it be labeled as outlier? If an object fits in a sparse or insignificant cluster, how can the labeling be performed within the cluster? Although clustering-based models have several drawbacks, they are good choices for most cases. Data stream is an interesting area for many researchers to apply cluster-based algorithms. For hierarchical- and partitioning-based clustering methods, speeding up the calculation process for large dataset and reducing CUP usage could be a suitable research direction. Detecting outliers

from lower density populations or within a low density cluster can make the algorithms robust.

7.2.4 Distance-based methods

Distance-based OD methods are popular in many application domains, the foundational technique behind this method is nearest neighbor model. A straightforward example of this method would be to apply KNN to a dataset, and based on distance of a data point, it's either reported as an outlier or non-outlier. By closely relating to density-based assumptions, distance-based methods have underlying assumptions that outlier points KNN distances are large compared to normal data points. In contrast, with clustering-based approach, they are more granular in their analytical procedure. Therefore these models are more effective in separating strong and weak outliers from malicious datasets. Again, referring to Fig. 7.1, it is evident that clustering-based methods face difficulties detecting outliers in noisy data. According to the definition of clustering-based outlier definition, outlier point "A" and nearest centroid of a cluster will be similar for both Figs. 7.1(a) and 7.1(b). On the contrary, distance-based methods consider distances from point "A," and noisy data are handled accordingly in terms of distance estimation. However, cluster-based methods can be modified to address the issue of noisy samples, in that case, these two methods have the same organization, as they are closely related. The distance-based algorithms provide scores to each datapoint incurring operational complexity proportional to $O(n^2)$. If binary labeling is expected as the outcome of the model, pruning techniques can be used to speed up the model substantially.

7.2.4.1 K-nearest neighbor models

KNN is one of the fundamental algorithms for distance-based OD approaches. Initially, nearest neighbor methods detect global outliers, and then assign them outlier scores. In KNN classification, distance information is investigated form a point to its neighbor, whether it's close or not. The fundamental idea is to utilize distance estimation to identify outliers. Knorr and Ng (1998) proposed a novel approach based on a non-parametric

technique that showed significant improvement over state-of-the-art OD algorithm at the time, especially for large dataset. Their approach differs from some of the previous method proposed by Yang et al. (2009b) and Satman (2013), where a user doesn't know about the underlying distribution of the dataset. Their computational complexity is $O\left(kN^2\right)$, where N is the number of the datasets and k is the dimensionality. In Knorr and Ng (1998), nested loop and indexed-based algorithm were applied to design OD models. Afterwards, Ramaswamy et al. (2000) proposed an improved technique that addressed the shortcomings of OD model by Knorr and Ng (1998), addressed computational cost, ranking method, and distance. They adopted the k*th* nearest neighbor that helps to ignore assigning distance parameter for the OD model. In another study, Knorr et al. (2000) expanded OD model proposed by Knorr and Ng (1998), modified nearest neighbor estimation by applying X-tree, KD-tree, R-tree, and indexing structure. For each example, the index structure is queried for nearest k points. Finally, top n number of outlier candidates are selected. However, the model falls apart when applied to large dataset of index structure.

Angiulli et al. (2006) proposed a technique that detects top-n number outliers from an unlabeled dataset. After that, the model predicts if a particular point is either an outlier or not. Top outliers get the highest weights; this is done by observing if a sample's calculated weight is higher than the top-n highest weights. Their approach incurs an $O(n^2)$ computational complexity. Later, Ghoting et al. (2008) developed an algorithm to address drawbacks of OD methods by Knorr and Ng (1998) and Ramaswamy et al. (2000), where they tried to improve the run time for high-dimensional feature space. They named the model recursive binning and re-projection (RBRP). In 2009, Zhang et al. (2009b) took a different path and projected an algorithm called local distance-based outlier factor (LDOF), which manages local outliers. Their study presented significant improvement compared to LOF (Breunig et al., 2000) in terms of range of neighbor size. This algorithm is similar in performance to KNN OD methods, such as COF (Tang et al., 2002). However, sensitivity on parameter value is insignificant. Later in 2013, a new model, called rank-based detection algorithm (RBDA), was proposed by Huang et al. (2013) to rank neighbors. It understands the meaning

and nature of high-dimensional dataset by providing a feasible solution. The key assumption of the model is this: objects will be similar and close to each other, thereby sharing similar neighborhood if they are generated from the same apparatus. Instead of taking object distance information from neighbors, the model considers individual objects ranks which are close to the degree of proximity of the object. Another method, proposed by Bhattacharya et al. (2015), applies reverse nearest neighbor and nearest neighbor as an extended study of RBDA.

Dang et al. (2015) applied an OD algorithm using KNN in large traffic data in big cities. The model they proposed detects outliers by exploiting the information among neighborhoods in which outliers are far from neighbors. This pipeline shows improved accuracy (95.5%), which is better than some statistical methods, such as GMM (80.9%) and KDE (95%). Despite improved accuracy, it has trouble keeping a single distance-based measure. Wang et al. (2015) used a least spanning tree to increase searching mechanism of neighbors of KNN algorithm. In another paper, Radovanović et al. (2015) proposed a reverse nearest technique to manage high-dimensional feature space. They presented the pipeline that can both manage low- and high-dimensional datasets. In terms of OD rates, this method works better than the original KNN method presented in Ramaswamy et al. (2000). Their method shows good performance on high-dimensional datasets. In contrast to OD model proposed by Ramaswamy et al. (2000), Jinlong et al. (2015) modified a technique to get the neighborhood information using a natural neighbor concept. In another study, Ha et al. (2015) implemented a heuristic technique to achieve k value by employing random iterative sampling. Recent study on OD in local KDE is investigated by Tang and He (2017). Several types of neighborhood information were examined by them, including k nearest, shared nearest, and reverse nearest neighbor. The KNN-based approaches are easy to implement despite their sensitivity to parameter selection and less superior performance.

7.2.4.2 Pruning techniques

Pruning technique is popular tool in ML models. A method, utilizing pruning technique method and randomization rule, based on nested loop, is presented by Bay and Schwabacher (2003). They modified the nested loop

technique, which was earlier known as quadratic $O(n^2)$ in performance and transformed into almost linear for most of the datasets. However, various assumptions in this pipeline resulted in poor performance. In another study, Angiulli and Fassetti (2007a) presented a generic pipeline, where outliers are detected by pushing data in an index. While developing the algorithm, they focused on minimizing input and output cost as well as CPU cost, because these costs were a major challenge in previous research (Knorr and Ng, 1998; Knorr et al., 2000; Ren et al., 2004a), where they achieved both demands simultaneously. Ren et al. (2004a) implemented a model to improvise the model proposed by Ramaswamy et al. (2000); they added pruning and labeling techniques to present a vertical distance-based OD algorithm. The method is implemented on both with and without pruning method, while adopting P-tree. Applying P-tree technique to other density-based OD can be a good future work. Later, another technique was developed to improvise OD model proposed by Ren et al. (2004a) for speeding up the detection process by Vu and Gopalkrishnan (2009), where similar pruning techniques are applied.

7.2.4.3 Time series data

Time series continuous data naturally create problems, such as uncertainty (Shukla et al., 2015), multidimensionality, notion of time, and concept drift, while applying them to an OD model. Usually, time series data are segmented by a time window. Two popular time series window methods are: a) sliding window (Angiulli and Fassetti, 2010), where two sliding endpoints are used to mark a window, and b) landmark window, where time points are identified to analyze *from-to* timeframes. A novel pipeline, proposed by Angiulli and Fassetti (2010), utilizes distance-based approach, where three different algorithms were developed for OD in time series data. They named the pipeline STORM (stream outlier miner). STORM utilizes two modules: data structure and stream manager, where the later collects continuous data streams, and the former is applied by the stream manager. However, sorting cost of window is a shortcoming of the algorithm, and colossal memory creates a burden, as it cannot fit properly into memory. Later, Lai et al. (2021a) performed OD time series benchmarked dataset and defined new context aware OD.

In another study, Yang et al. (2009a) developed several methods: Extra-N, Exact-N, Abstract-C, and Abstract-M to detect outliers based on neighborhood pattern information in the sliding window. This approach makes proper use of incremental OD by utilizing neighbor pattern in the sliding window of the dataset, which was not studied in earlier algorithms, such as DBSCAN (Zhang, 2013). This algorithm shows improved performance, linear memory utilization per object in a sliding window along with lower computational cost. Abstract-C applies a distance-based approach, whereas Extra-N, Exact-N, and Abstract-M utilize density-based cluster methods.

In another study, Angiulli and Fassetti (2007a), several issues were discussed in event detection, which were tackled by Kontaki et al. (2011), along with sliding window issues on time series data (Yang et al., 2009a). Angiulli and Fassetti (2007a) applied step function for processing the OD, wherein two algorithms parallelly utilize the sliding window. The primary focus in Kontaki et al. (2011) was to make the method flexible, lower storage usages, and enhance model efficiency. To support these ideas, three algorithms were proposed: COD, ACOD, and MCOD, short for continuous, advanced continuous and micro-cluster-based advanced OD, respectively. COD has two versions that support multiple values of k and a fixed radius R, where k and R are the parameters for OD algorithm. On the other hand, both multiple radius and k values are supported by ACOD. MCOD needs less distance calculation done for OD by minimizing query range. COD, compared to STORM and Abstract-C algorithm, reduces the number of objects in each window and requires less memory space. Another method was developed to process large data volume proposed by Cao et al. (2014b); it optimizes the range queries by not storing the objects in same window of same index structure. It is experimentally proven by the authors that MCOD is the most successful performing OD among COD, ACOD, and MCOD.

7.2.4.4 Advantages of distance-based methods

These methods don't rely on underlying distribution of data to detect outliers, thereby are straightforward algorithms. They also perform better compared to statistical-based methods and scale well for high-dimensional dataset because of their robust architecture.

7.2.4.5 Disadvantages of distance-based methods

Although distance-based methods perform better on high-dimensional feature spaces than statistical-based methods, the increasing dimensions issue reduces their performance. This is because different objects have distinctive attribution in the given dataset, which make it difficult for the model to measure distance among such objects. Also, if KNN is applied for computing distance-based OD, then the model becomes computationally expensive and unscalable. For data streams, distance-based methods face difficulties in both data distribution in local neighborhood and investigation of KNN in the time series data.

7.2.4.6 Research gaps and suggestions

Distance-based algorithm are effective mathematical tools to seek anomalies in a dataset. One major challenge is to scale for high-dimensional dataset (Aggarwal and Yu, 2001). Very large feature spaces and object's random attributions force models to underperform. Not only increasing feature space reduces the ability of the model to describe by distance measures, but also makes it difficult to comprehend the indexing approach to assigning neighbors. Additionally, multivariate data make the model less scalable when calculating distance measures. The models can be modified further by both improving execution time and memory usages. Another challenge is the quadratic complexity of the models, where researchers developed many techniques, including pruning and randomization (Bay and Schwabacher, 2003) and compact data structure (Bhaduri et al., 2011; van Hieu and Meesad, 2016). Distance-based methods are unable to detect local outliers, therefore often global information is calculated instead. To achieve desired scores from KNN algorithms, datasets need to be appropriate and properly processed. Selecting appropriate parameters, including proper k value, dictates performance of the model, and optimizing value of k and other parameters isn't easy always.

7.2.5 Ensemble methods

Recently, many domains, such as healthcare and technology, apply meta-algorithms for data mining problems, such as classification or clustering

to improve the solution. Such meta-algorithms create a series of multiple learning techniques: combinedly acts as a robust algorithm known as ensemble. Ensemble methods are mostly used in ML for their superior solutions compared to other traditional methods. These approaches are relatively new, and applied mostly on clustering and classification problems. The main idea behind this method is to train a dataset with multiple weak learners, while each learning outcome gets improved by a subsequent learner, therefore reducing the loss function. This working architecture lets the model be independent of dataset localizations. Although, detecting outliers using ensemble is not straightforward, many algorithms are proposed in recent years: bagging, boosting, bagged outlier representation ensemble (BORE), extreme gradient boosting OD (XGBOD), and isolation Forest (Lazarevic and Kumar, 2005; Rayana and Akoglu, 2016; Micenková et al., 2015; Zhao and Hryniewicki, 2019b; Liu et al., 2008). Bagging and boosting algorithms solve classification problems; for sequential methods XGBOD is applied; for hybrid and parallel models, BORE and isolation forest are applied.

One of the first ever ensemble method is known as bagging, refined recently by Lazarevic and Kumar (2005); it shows improved performance over large dimensional dataset by utilizing feature bagging techniques. This technique splits and creates random subsets of features and combines the outcome of multiple detection algorithms applied separately onto the subsets of features. Each algorithm is randomly assigned a small subset of feature to provide an outlier score; these scores are labeled to all the datapoints. They experimentally showed that bagging has improved performance, because it focuses on the outcome of multiple algorithms, where each algorithm targets a small portion of a feature.

In another study, an ensemble method is presented for outliers' detection by Aggarwal (2013), which was later discussed by many others (Kirner et al., 2017; Campos et al., 2018). Others proposed bagging (Lazarevic and Kumar, 2005) and boosting (Campos et al., 2018) from a classification context for ensemble analysis; also, alternative clustering (Müller et al., 2010) and multi view (Bickel and Scheffer, 2004) methods were proposed from a clustering context. Some critical questions were answered, such as how to

categorize if ensemble methods are independent or sequential, and how to categorize if ensemble methods are model- or data-centered? Ensemble algorithms are generally classified based on component independence. For instance, the components in boosting algorithms are not independent of each other, because results in each stage depend on prior executions, whereas bagging is the opposite, which makes their components independent of each other. Also, if the methods are model-centered, then the components of ensemble analysis are independent.

Later, several succeeding works have performed using ensembles for OD, including Nguyen et al. (2010), Kriegel et al. (2011), and Schubert et al. (2012), which face various challenges. One of the issues is to score comparison provided by mixture models and various functions for outliers and combine them to get a general outlier score. In another study, Schubert et al. (2012), based on outlier scores, compared the outlier ranking by observing similarity events. Their approach is a greedy technique that achieved good performance through differentiating actions. In another study, Nguyen et al. (2010) addressed problems with high-dimensional dataset and combined non-compatible OD method to form a unified approach. They implemented various scoring technique, each time to determine the degree of outlierness of a sample instead of using same approach repeatedly. Because of their heterogeneous approach, they called their method heterogeneous detector ensemble (HeDES), which represents combination of functions and heterogeneity affair. The HeDES, in contrast to methods proposed by Lazarevic and Kumar (2005), assign score types and scores for different outliers. The method shows improvement on real-world dataset. However, modification on the algorithm to handle large dimensional dataset can be a good research experiment.

Later in another study, Zimek et al. (2013) applied an arbitrary subsampling approach to calculate local density of nearest neighbors. When subsampling techniques are used on a dataset, usually training objects can be obtained without replacement, therefore they enhance OD performance. Also, subsampling technique with other OD can give good results as well. Zimek et al. (2014a), later investigated an ensemble learning approach for OD; the pipeline brings a perturbation technique to account for different

diversities in different outlier detectors as well as adopting a method that considers outlier rankings combinedly and distinctively.

As we suggested earlier, Pasillas-Díaz and Ratté (2016) did apply both feature bagging and subsampling technique together. Each technique is assigned to a different task: feature bagging extracts various information during each iteration, whereas subsampling technique scores different sets of data. However, getting variance of objects by using feature bagging was a drawback and the result depends on the size of the subsample. Except for these shortcomings, the method has improved in performance. Another method that dynamically combines the score values, an unsupervised framework, is proposed by Zhao and Hryniewicki (2019a); they developed a way to combine and select outlier scores, even if the ground truth is absent. Zhao et al. (2018) proposed a similar approach as Zhao and Hryniewicki (2019a), and implemented four variations of it.

7.2.5.1 Advantages of ensemble methods

The ensemble analysis is better for investigating outliers because of their much better prediction models. Bagging and boosting are two popular and efficient algorithms. They are robust and less dependent on a particular dataset in data mining processes. Ensemble methods are suitable for adopting high-dimensional datasets, which used to be a burden for traditional OD algorithms.

7.2.5.2 Disadvantages of ensemble methods

Mathematically, ensemble analysis isn't that much robust as other data mining techniques, it is because they are not properly developed yet. This results in poor feature evaluation along with difficulties in selecting contextual meta-detectors. Various algorithms are combinedly working, and since the sample space is smaller, researchers face challenges managing real data in some cases using these methods.

7.2.5.3 Research gaps and suggestions

Although ensemble analysis has shown robust results, there are still issues that need to be fixed. They show good performance when streaming data has noise in it, because individual classifiers face difficulties when it

comes to the quality of data and processing time. However, combinedly, those classifiers yield good outcome. Zimek et al. (2014b) addressed multiple challenges along with data quality and processing time, which has been brought under consideration by developing models, such as Nguyen et al. (2010), Aggarwal and Sathe (2015), Liu et al. (2008), and Kriegel et al. (2011), to improve ensemble analysis for detecting outliers. Also, several research gaps have been addressed by Zimek et al. (2014b), although ranking outliers from different detectors and diversifying principal proposals remains an open research challenge. Several techniques (Zimek et al., 2014b; Rayana and Akoglu, 2016) don't require detector selection process, therefore these methods, in absence of detector selection process, hardly help in speeding up identifying unknown outliers.

7.2.6 Learning-based methods

Learning-based methods are applied to different sub-discipline in ML. In this section, we discuss four categories: Subspace, Active, Graph-based and Deep Learning (DL).

7.2.6.1 Subspace learning models

OD models that have been discussed so far, usually identifies outliers from all the space and dimension. However, outliers often represent different attributes in the local neighborhood on declining dimensional subspace. To address this issue, Zimek et al. (2013) presented that appropriate selection of a subset carries significant attribute information. On the contrary, residual attributes have less importance or sometime has no importance at all, and they delay the OD process. Subspace learning in OD is popular for high-dimensional areas. The fundamental focus is to identify dissimilar dimension subsets and meaningful outliers form a given data. We can further categorize these studies into two subcategories: relevant subspace methods (Huang et al., 2013; Muller et al., 2008) and sparse subspace methods (Zhang et al., 2009a; Dutta et al., 2016). The sparse subspace learning techniques project high-dimensional datasets onto sparse and low-dimensional subspace. The outliers are the ones located in sparse subspace, because they are characterized as lower density. Projecting high-

dimensional space onto sparse subspace is time consuming, therefore a big challenge. Aggarwal and Yu (2005) addressed this issue and proposed a method for effective subspace exploration, where an evolutionary algorithm gathers the subspaces. Here, initial population dictates the algorithms performance evaluation.

Later, Zhang et al. (2009a) proposed a method that focuses on spares subspace technique's path. The method applies the idea of lattice to denote subspace relationship; sparse subspace is related to lower density coefficient. Applying the idea of lattice makes the model perform poorly and complex in architecture. A new way to get sparse space is implemented by Dutta et al. (2016), here sparse encoding is used to transform objects to multiple linear space. Relevant subspaces are used by outlier detectors to find local information as they are essential features in this case. A relevant subspace method is proposed by Huang et al. (2013), called subspace OD (SOD). The method examines correlation of every object with its shared nearest neighbor; instead of taking distance from objects to its neighbors, the model considers ranks of each object that is close to the proximity of the object. Here, primarily the variance of the features is focused by SOD. Another method, in contrast to SOD, signifies the relationship between features is proposed by Müller et al. (2011).

In another but similar study, Kriegel et al. (2009c) presented OD method that achieve relevant subspace, where distances are computed by Mahalanobis technique through gamma distribution. Principal component analysis is used in this context. In contrast to Müller et al. (2011), the key difference is the requirement of large local dataset to recognize the abnormality trend. This impacts the scalability and flexibility of the method in a gradual manner. To tackle flexibility problem, a similar method is proposed by Keller et al. (2012) that identifies subspaces and ranks the outliers. The Monte Carlo method, a sampling technique, is implemented, called high contrast subspace (HiCS), where LOF scores are combined based on HiCS values. In another study by van Stein et al. (2016), after achieving HiCS instead of using LOF scores, LoOP scores are used to calculate the degree of outlierness.

Nevertheless, though the subspace learning methods are highly efficient, for OD, in several cases, they are computationally expensive. Searching for subspaces in high-dimensional space is a daunting task, which makes the pipeline more complex.

7.2.6.2 Active learning models

Active learning methods are semi-supervised learners through input sources or by interacting with users to get the desired outputs (Das et al., 2016). For instance, for large dataset that require labeling, doing so manually is an exhaustive process. Since the method queries the user iteratively, this supervised approach is called active learning. When an active learning algorithm is trained, it can find smaller portions of the dataset that contain the labels. This helps the algorithm to re-train and boost for improvements. Also, by querying labels for instances from the user iteratively, it provides better suggestions. Recently, researchers have been focusing on this approach for OD in different domains (Zhang et al., 2009a; Dutta et al., 2016; Yiyong et al., 2007; Muller et al., 2008). Aggarwal and Yu (2005) applied active learning to unveil the reason for flagging the outliers and the reason behind high computational demand for estimating density for OD methods. The sampling process that was applied is called ensemble active learning. Later, Görnitz et al. (2014) applied an active learning method for OD; they alternatively repeated the learning process and updated the model to improve prediction results. After training on improved and unlabeled examples, the active learning method is applied.

In another study, input from a human analyst is provided to get better result using active learning (Dutta et al., 2016; Yiyong et al., 2007). Although they selected good portion of instances for the querying process, they didn't provide any explanation or clear insight or interpretation for the model design procedure. However, later they attempted to address the issues; a modified active learning approach is proposed by Das et al. (2019). They called the method glocalized anomaly detection (GLAD). Their primary focus is to adopt ensemble outlier detectors so that they can solve active learning problems. The end users have the control to global outlier detector; GLOD attains the local weights of data instance by learning automatically. Here,

label feedback helps to implement this process. Also, proper tuning of ensemble detectors helps to identify maximum number of accurate outliers. This pipeline is also known as human-in-the-loop, where label feedback is achieved by a human analyst in each iteration round. In another study, Zha et al. (2020) proposed a deep reinforcement learning-based OD algorithm, which detects outliers by achieving balance between long- and short-term rewarding processes.

Even though active learning serves a great purpose in OD community, there is still scope for improvement. Receiving inputs from human analyst is a daunting task; an AI assurance method is required to minimize the effect of false positive labeling, while designing the model. Active learning methods are better at identifying outliers. However, more interpretation techniques should be adopted to explain the results.

7.2.6.3 Graph-based learning models

Graphs are known as data structure that can adapt various algorithm, especially neural network, to perform learning task, such as clustering, classification, and regression. Applications of graph data are getting popular for OD in various sectors. Initially, these algorithms transform each vector node into a real vector. Then the outcome is a vector representation of each node, where information gets preserved in the graph. After achieving a real vector, one can apply it to a neural network.

Many algorithms have proposed especially OD in graph data; a broad review of graph-based OD approaches are presented by Akoglu et al. (2014) and Ma et al. (2021). The authors have presented state-of-the-art techniques and several research challenges. They also discussed the importance of using graph-based OD, where graph-based approach shows the interdependency state of the data, robust, and insightful distribution. A very first graph-based detection framework, called "*Outrank*," is proposed by Moonesinghe and Tan (2008). They established entirely undirected graphs using the original dataset and a technique is applied to the predefined graph, called Markov random walk. Markov random walk stationary distribution values are used to score all samples. Later, a novel approach is presented by Wang et al. (2018a), where objects' local information together with combined representation of the graph is adopted. They addressed the

issue of false positive rates in OD, where graph-based method ignores local information of an object around each node. Therefore local information of each object's surrounding of each node is collected, which helps to construct the graph. Thereafter, outlier scores are provided by randomly "walking through" the graph. This method adopts multiple neighborhood graphs, where outlier scores are generated by walking through predefined graph. The authors conclude that their model showed good improvement. The graph-based OD methods are relatively new and promising technique, having great potentials for OD in many domains.

7.2.6.4 Deep learning models

Deep Learning (DL) methods are a member of the ML family that are mainly applied for representation and patterns learning by incorporating artificial neural networks (ANN). Application of DL can be supervised, unsupervised or semi-supervised. These methods are getting popular because of their high accuracy on detecting outliers in critical infrastructure, healthcare, and defense (amongst many other domains). A survey in DL presented by Chalapathy and Chawla (2019) reviewed multiple DL-based OD techniques and their evaluation. These models are effective for large dimensional dataset and can understand hierarchical information on features. Additionally, they are better for separating the boundary conditions between normal and abnormal behavior in time series dataset. Supervised DL models explore outliers by training and classifying the relationship between features and labels. For example, supervised models, such as multiclass classifier, are used to detect fraudulent transaction in healthcare (Chalapathy and Chawla, 2019). Although, supervised models provide great results, unsupervised and semi-supervised models are mostly utilized. This is because, supervised models require labeling for each sample, so it's a daunting task to label each sample. Therefore unsupervised and semi-supervised models are a better selection in real-world application with big datasets.

Semi-supervised DL methods for OD is the most appealing approach, given it provides flexibility regarding labeling requirements. The models use normal instances as references to identify outliers. Deep autoencoder, a semi-supervised deep neural learning model that can be applied to a dataset to find outliers. If enough training sets with normal events can be

provided, the autoencoder can understand the inter-dependency of features. It generates a *reconstruction error* for all input features by encoding and decoding them, where the abnormal instances have higher reconstruction error.

Unsupervised DL OD techniques focus on essential features to find outliers form dataset. They label the dataset, which is initially not labeled. The autoencoder is a popular unsupervised DL OD technique (Chen et al., 2017). In recent research (Zhou and Paffenroth, 2017; Chalapathy et al., 2017), unsupervised DL OD algorithm shows great effectiveness. Unsupervised models can be divided into two subcategories, such as model architecture embracing hybrid models (Erfani et al., 2016) and autoencoders (Andrews et al., 2016). The autoencoder-related models measures the degree of outlierness by observing reconstruction error of each feature space through adopting the value of residual vector. Hendrycks et al. (2018) implemented an approach for improving the OD technique called outlier exposure. They identified a classification model by performing iteration to understand the heuristics; it helps to distinguish between distributed samples and outliers.

A universal framework that utilizes DL technique to log online OD and analysis, called Deeplog, was presented by Du et al. (2017). To model the system log, Deeplog applies long short-term memory architecture. The algorithm learns and encodes the whole logging process. In contrast to other methods where outliers are detected in each session, Deeplog learns outliers for every log entered. In high-performance computing system, Borghesi et al. (2018) developed OD technique using autoencoder (Neural Network). A set of autoencoders are trained with the supercomputer nodes to learn the normal behavior, afterwards those autoencoders can identify abnormal behaviors.

Based on training mechanism, deep leaning OD methods can engage either one class neural network or deep hybrid models (Chalapathy and Chawla, 2019). Adopting deep neural networks, deep hybrid models mainly emphasize on extracting feature from the autoencoder after learning the hidden representation from the autoencoders. Most OD algorithms use them as inputs, such as one class SVM. Because of the shortage of labeled

datasets for OD, hybrid approaches have notable limitations, despite their performance maximization for OD. Therefore features that are rich and differentiable are applicable for deep hybrid models. To address and solve this problem, Ruff et al. (2018) introduced deep one class classification, and Chalapathy et al. (2018) introduced one class neural network.

7.2.6.5 Advantages of learning-based methods

In graph-based approach, interdependency of datapoints gets revealed by exhibiting an intuitive representation for OD. DL methods, however, are good for investigating the hierarchical discrimination between features in each dataset. Also, they have improved performance on large dimensional time series data. For time series data, they have effective ways to set boundaries between normal and outlier data.

7.2.6.6 Disadvantages of learning-based methods

Learning-based model, especially subspace learning is computationally expensive. Generally, not all traditional DL methods are good on increasingly large amount of feature spaces, therefore detection of outliers could become more challenging.

7.2.6.7 Research gaps and suggestions

Not all methods in neural network can effectively differentiate the boundary between normal and outlier points, which is a vital task for data mining. Moreover, further research is required for recurrent neural networks, long short-term memory, deep believe network for OD. Kwon et al. (2019) and Chalapathy and Chawla (2019) are surveys on deep neural network OD that present further insights.

7.3 Tools for outlier detection

There are many of the shelf libraries and tools available to apply OD research and development. Among many tools, we include the most popular ones that are frequently used by the research community:

a) Scikit-learn (Python): Scikit learn is a well-known tool for AI research. This tool has some popular algorithms, including isolation forest (Liu et al., 2008) and local outlier factor (Breunig et al., 2000) etc.

b) Python outlier detection (PyOD) (Python): PyOD is another popular tool for OD in multivariate data. This library is widely used in academic research and some commercial purposes; it includes ensemble methods and several DL techniques (Ramakrishnan et al., 2019; Kalaycı and Ercan, 2018).

c) ELKI (Java): ELKI, which stands for environment for developing KDD-applications supported by index-structures, is a Java-based open-source platform for developing KDD applications and other data mining OD algorithms. The source code is written in Java, it provides benchmarking and simple fairness assessment test for the algorithms (Achtert et al., 2010).

d) Python streaming anomaly detection (PySAD) (Python): PySAD is an open-source Python-based library for streaming data to identify outliers. It contains a collection of algorithms, including more than 15 online detector algorithm and two PyOD detectors setting for data (Yilmaz and Kozat, 2020).

e) Scalable unsupervised OD (SUOD) (Python): SUOD works on top of PyOD; it's an unsupervised learning OD acceleration framework for large-scale dataset training and predictions (Zhao et al., 2020).

f) Rapid miner (Java): Rapid miner (Kalaycı and Ercan, 2018) is a Java-based OD extension. It adopts unsupervised approach, including COF (Tang et al., 2002), LOF (Breunig et al., 2000), LOCI (Papadimitriou et al., 2003), LoOP (Kriegel et al., 2009b).

g) MATLAB®: MATLAB is a user-friendly commercial software that supports many OD algorithms.

h) Time-series outlier detection system (TODS) (Python): It's a python based full-stack environment for detecting outliers in multivariate data streams (Lai et al., 2020).

i) Skyline (Python): Skyline detects anomalies in near real-time.

j) Telemanom (Python): Telemanom adopts long short-term memory architecture for multivariate time series data to detect outliers.

k) DeepADoTS (Python): A collection of DL benchmarking pipelines for OD for time series data.
l) Numerical anomaly benchmark (NAB) (Python): For real-time and streaming data; NAB is used to evaluate multiple algorithms for benchmarking purpose.
m) Datastream.io (Python): Datastream.io is an open-source tool for detecting outliers in real time data.

7.4 Datasets for outlier detection

In data mining problems, two types of data are used to train any OD models, including real data and synthetic data. Real data are expensive to generate and distribute because of their security and commercial aspect. In this chapter, we enlist multiple real datasets to begin modeling OD problems. Some of the most popular OD datasets are as following:

a) University of California Irvine (UCI) repository: The UCI repository (https://archive.ics.uci.edu/ml) provides more than hundreds of datasets; many researchers use these datasets for evaluating their algorithm. However, this server mostly contains dataset for classification algorithms.
b) ELKI dataset: (http://elki-project.github.io/datasets/outlier): ELKI has numerous
available datasets that can be used for different type of OD algorithm and for assessing model parameters.
c) Outlier detection datasets (ODDS) (http://odds.cs.stonybrook.edu/#table1): ODDS contain various types of datasets and they constitute a good source for training-testing OD algorithms. Some of the popular datasets from this server are time series multivariate and univariate datasets, high-dimensional data, and time series graph data.
d) Anomaly detection meta-analysis benchmarks (http://ir.library.oregonstate.edu/concern/datasets/47429f155): Oregon State University has enriched datasets for evaluating various OD algorithm.

e) Harvard database: (dataverse.harvard.edu/dataset): This server contains datasets that can be used for benchmarking unsupervised algorithm. It also contains several datasets for supervised OD models.

f) Skoltech anomaly benchmark (SKAB) (http://github.com/waico/skab): This repository contains approximately 34 datasets; authorities plan to add more than 300 datasets in the near future for collective anomalies and point anomalies.

All the above-mentioned sources provide many collective datasets to begin with OD studies. However, most of the real-world datasets are not available publicly, because of security and privacy concerns. For instance, data from critical infrastructure, such as electricity transmission, water distribution, and healthcare aren't available publicly. Therefore synthetic data are an alternative and next best option for creating specific domain-related models. For example, BATADAL (http://batadal.net/data.html) presents a synthetic data by creating virtual supervisory control and data acquisition system (SCADA) on top of a water distribution system network (Daneels and Salter, 1999). Since most real SCADA data aren't publicly available, this synthetic dataset is a good choice for researchers. In data mining problems, various evaluation techniques are implemented for the OD algorithms to measure "goodness." These evaluation techniques focus on OD rates and run times of the algorithm. Mostly adopted evaluation measurements are Precision, R-Precision, Area Under the Curve (AUC), Average Precision, Receiver Operating Characteristics (ROC), Correlation Coefficient, and Rank Power (RP) (Domingues et al., 2018).

7.5 AI assurance and outlier detection

In this chapter, we discuss several working algorithms for OD in data mining problems for AI assurance. According to (Batarseh et al., 2021), AI assurance can be defined as:

"A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users."

Additionally, in their review paper, the authors added ten metric scoring schemes to present a systematic comparison among existing AI assurance approaches. To verify an AI model, six assurance goals need to be verified for an AI system: fairness, trustworthiness, ethics, safety, security, and explainability. The authors address the complexity of recent AI algorithms and the necessity of investigating algorithm variance, bias, clarity, and awareness to measure these AI assurance goals. In general, AI assurance goals can be achieved by either model specific or model agnostic approach. Model specific approaches target specific AI algorithms for quantifying or validating assurance goals, whereas model agnostic approaches are generic and have universal frameworks that can verify all AI algorithms for assurance goals. Despite the challenges, AI assurance is necessary. OD is at the heart of assurance, as it improves the overall quality of the data.

Data quality needs to be assured as well. If the underlying data is invalid, then AI algorithms will have undesirable outcomes. OD algorithms measure two important aspects of data assurance: safety and security. This is because, analyzing a dataset for outlier not only means investigating abnormal samples, but also represents faults or intrusions in the system by adversaries. For instance, ANN autoencoders detect outliers using a reconstruction error, where the errors are generated during encoding and decoding process of a dataset. Higher reconstruction errors are an indication of an object being an outlier or an attack on the system. Therefore reconstruction errors, in this context, can be considered as safety and security measure. Other data assurance goals can be achieved depending on the context of application domain and AI algorithm used.

Assurance goals, especially fairness and ethics, can be achieved by removing bias in the dataset. However big data generated by real-world source almost always have bias (Verma et al., 2021). Some of the most common data biases are activity bias, selection bias, bias due to system drift, omitted variable bias, and societal bias. For identifying the reason behind any bias, one should investigate how the data are generated. Most common practice of data bias identification is to perform Exploratory Data Analysis (EDA) (Tukey, 2020). In a recent study, Amini et al. (2019) presented a debiasing technique during post processing after training with AI algorithm.

Their method adopts DL-based model to understand the latent data distribution during training stage in an unsupervised manner, thereby making the approach robust for debiasing. In another study, Bolukbasi et al. (2016) showed a debiasing technique to mitigate gender bias. For model assurance, in a recent study, Shekhar et al. (2020) applied a novel framework based on deep-autoencoder for fairness called fairness-aware OD (FairOD). They focused on formalizing the definition of fair OD algorithm with desirable properties. Data bias can yield unfair, unethical, and untrustworthy decisions by AI algorithms, therefore bias needs to be identified before training the AI model. Data bias can be also detected using OD algorithms.

7.6 Conclusions

This chapter reviews the state-of-the-art in approaches for outlier analysis. We group OD methods into several categories: Distance, Statistical, Density, Clustering, Learning and Ensemble-based methods. For each category, we present relevant algorithms, their significant importance, and drawbacks.

For distance-based methods, especially ones that use KNN based models, are sensitive to the parameter selection process, including the value of k. Therefore an appropriate k parameter selection is important for the models that rank neighbors for OD. Clustering-based methods generally are not explicitly suitable as they were not designed to facilitate OD. However ensemble methods that combine results from a collection of dissimilar detectors provide much improved outcomes. Ensemble methods have lower execution time, but high-quality OD results. Regarding model evaluation, effectively assessing an OD algorithm is still an open research challenge. Also, in many cases, it's a daunting task to evaluate a model when a ground truth is absent and outliers aren't that frequent. Deep neural network-based OD models are gradually becoming popular because of their effective measures and quality results. ANN-based autoencoders can detect outliers, even if sensor network data are compromised and concealed by an adversarial attack. Nonetheless, DL-based models are advanced and difficult to design. Moreover, enough investigations are required to unlock the full potential of DL-based models for detecting outliers in real-world applications. Lastly,

an important notion to note, OD models need to be assured, because AI algorithms ought to be safe and secure from unwanted outliers.

References

Abid, A., Kachouri, A., Mahfoudhi, A., 2017. Outlier detection for wireless sensor networks using density-based clustering approach. IET Wireless Sensor Systems 7, 83–90. https://doi.org/10.1049/iet-wss.2016.0044.

Achtert, E., Kriegel, H.-P., Reichert, L., Schubert, E., Wojdanowski, R., Zimek, A., 2010. Visual evaluation of outlier detection models. In: Kitagawa, H., Ishikawa, Y., Li, Q., Watanabe, C. (Eds.), Database Systems for Advanced Applications. Springer Berlin Heidelberg, pp. 396–399.

Aggarwal, C., 2017. Outlier Analysis.

Aggarwal, C.C., 2013. Outlier ensembles: position paper. SIGKDD Explorations Newsletter 14 (2), 49–58. https://doi.org/10.1145/2481244.2481252.

Aggarwal, C.C., 2016. Outlier Analysis, 2nd ed. Springer, New York, NY, USA.

Aggarwal, C.C., Han, J., Wang, J., Yu, P.S., 2003. A framework for clustering evolving data streams. In: Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29, pp. 81–92.

Aggarwal, C.C., Han, J., Wang, J., Yu, P.S., 2004. A framework for projected clustering of high dimensional data streams. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, pp. 852–863.

Aggarwal, C.C., Sathe, S., 2015. Theoretical foundations and algorithms for outlier ensembles. SIGKDD Explorations Newsletter 17 (1), 24–47. https://doi.org/10.1145/2830544.2830549.

Aggarwal, C.C., Yu, P.S., 2001. Outlier detection for high dimensional data. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, pp. 37–46.

Aggarwal, C.C., Yu, P.S., 2005. An effective and efficient algorithm for high-dimensional outlier detection. The VLDB Journal 14 (2), 211–221. https://doi.org/10.1007/s00778-004-0125-5.

Aggarwal, C.C., Zhao, Y., Yu, P.S., 2011. Outlier detection in graph streams. In: 2011 IEEE 27th International Conference on Data Engineering, pp. 399–409.

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp. 94–105.

Akoglu, L., Tong, H., Koutra, D., 2014. Graph-based anomaly detection and description: a survey. CoRR, arXiv:1404.4679.

Al-Zoubi, M.B., 2009. An effective clustering-based approach for outlier detection. European Journal of Scientific Research 28, 310–316.

Alrawashdeh, K., Purdy, C., 2016. Toward an online anomaly intrusion detection system based on deep learning. In: 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 195–200.

Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D., 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 289–295.

Andrews, J., Morton, E., Griffin, L., 2016. Detecting anomalous data using auto-encoders. International Journal of Machine Learning and Computing 6, 21.
Angiulli, F., Basta, S., Pizzuti, C., 2006. Distance-based detection and prediction of outliers. IEEE Transactions on Knowledge and Data Engineering 18 (2), 145–160. https://doi.org/10.1109/TKDE.2006.29.
Angiulli, F., Fassetti, F., 2007a. Detecting distance-based outliers in streams of data. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 811–820.
Angiulli, F., Fassetti, F., 2007b. Very efficient mining of distance-based outliers. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 791–800.
Angiulli, F., Fassetti, F., 2010. Distance-based outlier queries in data streams: the novel task and algorithms. Data Mining and Knowledge Discovery 20 (2), 290–324. https://doi.org/10.1007/s10618-009-0159-9.
Assent, I., Kranen, P., Baldauf, C., Seidl, T., 2012. AnyOut: anytime outlier detection on streaming data. In: Lee, S., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (Eds.), Database Systems for Advanced Applications. Springer Berlin Heidelberg, pp. 228–242.
Bai, M., Wang, X., Xin, J., Wang, G., 2016. An efficient algorithm for distributed density-based outlier detection on big data. Neurocomputing 181, 19–28. https://doi.org/10.1016/j.neucom.2015.05.135.
Batarseh, F.A., Freeman, L., Huang, C.-H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 60. https://doi.org/10.1186/s40537-021-00445-7.
Bay, S.D., Schwabacher, M., 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 29–38.
Bhaduri, K., Matthews, B.L., Giannella, C.R., 2011. Algorithms for speeding up distance-based outlier detection. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 859–867.
Bhattacharya, G., Ghosh, K., Chowdhury, A., 2015. Outlier detection using neighborhood rank difference. Pattern Recognition Letters, 60–61. https://doi.org/10.1016/j.patrec.2015.04.004.
Bhosale, S.P., 2014. A Survey: Outlier Detection in Streaming Data Using Clustering Approached.
Bickel, S., Scheffer, T., 2004. Multi-view clustering. In: Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004, pp. 19–26.
Boedihardjo, A., Lu, C.-T., Chen, F., 2013. Fast adaptive kernel density estimator for data streams. Knowledge and Information Systems 42. https://doi.org/10.1007/s10115-013-0712-0.
Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. CoRR, arXiv:1607.06520.
Bordogna, J.T., Brown, D.E., Conklin, J.H., 2007. Design and implementation of an automated anomaly detection system for crime. In: 2007 IEEE Systems and Information Engineering Design Symposium, pp. 1–6.

Borghesi, A., Bartolini, A., Lombardi, M., Milano, M., Benini, L., 2018. Anomaly detection using autoencoders in high performance computing systems. CoRR, arXiv: 1811.05269.
Braei, M., Wagner, S., 2020. Anomaly detection in univariate time-series: a survey on the state-of-the-art. CoRR, arXiv:2004.00433.
Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. SIGMOD Record 29 (2), 93–104. https://doi.org/10.1145/335191.335388.
Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J., 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data 10 (1). https://doi.org/10.1145/2733381.
Campos, G., Zimek, A., Meira Jr, W., 2018. An Unsupervised Boosting Strategy for Outlier Detection Ensembles, pp. 564–576.
Cao, F., Ester, M., Qian, W., Zhou, A., 2006. Density-Based Clustering over an Evolving Data Stream with Noise.
Cao, K., Shi, L., Wang, G., Han, D., Bai, M., 2014a. Density-based local outlier detection on uncertain data. In: Li, F., Li, G., Hwang, S., Yao, B., Zhang, Z. (Eds.), Web-Age Information Management. Springer International Publishing, pp. 67–71.
Cao, L., Yang, D., Wang, Q., Yu, Y., Wang, J., Rundensteiner, E.A., 2014b. Scalable distance-based outlier detection over high-volume data streams. In: 2014 IEEE 30th International Conference on Data Engineering, pp. 76–87.
Cateni, S., 2008. Outlier detection methods for industrial applications. In: Colla, V. (Ed.), Rijeka: IntechOpen (Ch. 14).
Chalapathy, R., Chawla, S., 2019. Deep learning for anomaly detection: a survey. CoRR, arXiv:1901.03407.
Chalapathy, R., Menon, A.K., Chawla, S., 2017. Robust, deep and inductive anomaly detection. CoRR, arXiv:1704.06743.
Chalapathy, R., Menon, A.K., Chawla, S., 2018. Anomaly detection using one-class neural networks. CoRR, arXiv:1802.06360.
Chen, J., Sathe, S., Aggarwal, C., Turaga, D., 2017. Outlier detection with autoencoder ensembles. In: Proceedings of the 2017 SIAM International Conference on Data Mining (SDM), pp. 90–98.
Chen, Y., Tu, L., 2007. Density-based clustering for real-time stream data. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142.
Chenaghlou, M., Moshtaghi, M., Leckie, C., Salehi, M., 2017. An efficient method for anomaly detection in non-stationary data streams. In: GLOBECOM 2017 - 2017 IEEE Global Communications Conference, pp. 1–6.
Chenaghlou, M., Moshtaghi, M., Leckie, C., Salehi, M., 2018. Online Clustering for Evolving Data Streams with Online Anomaly Detection, pp. 508–521.
Chenaoua, K.S., Kurugollu, F., Bouridane, A., 2014. Data cleaning and outlier removal: application in human skin detection. In: Paper Presented at 5th European Workshop on Visual Information Processing (EUVIP).
Dalatu, P., Fitrianto, A., Mustapha, A., 2017. A Comparative Study of Linear and Nonlinear Regression Models for Outlier Detection.
Daneels, A., Salter, W., 1999. What Is SCADA.

Dang, T.T., Ngan, H.Y.T., Liu, W., 2015. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In: 2015 IEEE International Conference on Digital Signal Processing (DSP), pp. 507–510.

Das, S., Islam, M.R., Jayakodi, N.K., Doppa, J.R., 2019. Active anomaly detection via ensembles: insights, algorithms, and interpretability. CoRR, arXiv:1901.08930.

Das, S., Wong, W.-K., Dietterich, T., Fern, A., Emmott, A., 2016. Incorporating expert feedback into active anomaly discovery. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 853–858.

Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J., 2018. A comparative evaluation of outlier detection algorithms: experiments and analyses. Pattern Recognition 74.

Du, H., Zhao, S., Zhang, D., 2015. Robust local outlier detection. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 116–123.

Du, M., Li, F., Zheng, G., Srikumar, V., 2017. DeepLog: anomaly detection and diagnosis from system logs through deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1285–1298.

D'Urso, C., 2016. EXPERIENCE: glitches in databases, how to ensure data quality by outlier detection techniques. Journal of Data and Information Quality 7 (3). https://doi.org/10.1145/2950109.

Dutta, J.K., Banerjee, B., Reddy, C.K., 2016. RODS: rarity based outlier detection in a sparse coding framework. IEEE Transactions on Knowledge and Data Engineering 28 (2), 483–495. https://doi.org/10.1109/TKDE.2015.2475748.

Edgeworth, F.Y., 1887. XLI. On discordant observations. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 23 (143), 364–375. https://doi.org/10.1080/14786448708628471.

Elahi, M., Li, K., Nisar, W., Lv, X., Wang, H., 2008. Efficient clustering-based outlier detection algorithm for dynamic data stream. In: 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5, pp. 298–304.

Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition 58, 121–134. https://doi.org/10.1016/j.patcog.2016.03.028.

Eskin, E., 2000. Anomaly Detection over Noisy Data Using Learned Probability Distributions. ICML.

Ester, M., Kriegel, H., Sander, J., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD.

Fan, H., Zaïane, O.R., Foss, A., Wu, J., 2009. Resolution-based outlier factor: detecting the top-n most outlying data points in engineering data. Knowledge and Information Systems 19 (1), 31–51. https://doi.org/10.1007/s10115-008-0145-3.

Feng, H., Liang, L., Lei, H., 2017. Distributed outlier detection algorithm based on credibility feedback in wireless sensor networks. IET Communications 11, 1291–1296. https://doi.org/10.1049/iet-com.2016.0986.

Gao, J., Hu, W., Zhang, Z. (Mark), Zhang, X., Wu, O., 2011. RKOF: robust kernel-based local outlier detection. In: Huang, J.Z., Cao, L., Srivastava, J. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 270–283.

Gebhardt, J., Goldstein, M., Shafait, F., Dengel, A., 2013. Document authentication using printing technique features and unsupervised anomaly detection. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 479–483.

Gebremeskel, G.B., Yi, C., He, Z., Haile, D., 2016. Combined data mining techniques based patient data outlier detection for healthcare safety. International Journal of Intelligent Computing and Cybernetics 9 (1), 42–68. https://doi.org/10.1108/IJICC-07-2015-0024.

Ghanbari, S., Hashemi, A.B., Amza, C., 2014. Stage-aware anomaly detection through tracking log points. In: Proceedings of the 15th International Middleware Conference, pp. 253–264.

Ghoting, A., Parthasarathy, S., Otey, M., 2008. Fast mining of distance-based outliers in high-dimensional datasets. Data Mining and Knowledge Discovery 16, 349–364. https://doi.org/10.1007/s10618-008-0093-2.

Goldstein, M., Dengel, A., 2012. Histogram-Based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm.

Goldstein, M., Uchida, S., 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE 11 (4), 1–31. https://doi.org/10.1371/journal.pone.0152173.

Görnitz, N., Kloft, M., Rieck, K., Brefeld, U., 2014. Toward supervised anomaly detection. CoRR, arXiv:1401.6424.

Guha, S., Rastogi, R., Shim, K., 2001. Cure: an efficient clustering algorithm for large databases. Information Systems 26 (1), 35–58. https://doi.org/10.1016/S0306-4379(01)00008-4.

Gupta, M., Gao, J., Aggarwal, C.C., Han, J., 2014. Outlier detection for temporal data: a survey. IEEE Transactions on Knowledge and Data Engineering 26 (9), 2250–2267. https://doi.org/10.1109/TKDE.2013.184.

Ha, J., Seok, S., Lee, J.-S., 2015. A precise ranking method for outlier detection. Information Sciences 324. https://doi.org/10.1016/j.ins.2015.06.030.

Hadi, S., Imon, R., Werner, M., 2009. Detection of outliers. WIREs: Computational Statistics 1 (1), 57–70. https://doi.org/10.1002/wics.6.

Hawkins, D., 1980. Identification of Outliers. Springer, Netherlands.

Hendrycks, D., Mazeika, M., Dietterich, T.G., 2018. Deep anomaly detection with outlier exposure. CoRR, arXiv:1812.04606.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., Kanamori, T., 2011. Statistical outlier detection using direct density ratio estimation. Knowledge and Information Systems 26, 309–336. https://doi.org/10.1007/s10115-010-0283-2.

Hinneburg, A., Keim, D., 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD.

Huang, H., Mehrotra, K., Mohan, C.K., 2013. Rank-based outlier detection. Journal of Statistical Computation and Simulation 83 (3), 518–531. https://doi.org/10.1080/00949655.2011.621124.

Iglesias Vázquez, F., Zseby, T., Zimek, A., 2018. Outlier Detection Based on Low Density Models, pp. 970–979.

Jin, W., Tung, A.K.H., Han, J., Wang, W., 2006. Ranking outliers using symmetric neighborhood relationship. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (Eds.),

Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 577–593.

Jinlong, H., Zhu, Q., Yang, L., Feng, Ji., 2015. A non-parameter outlier detection algorithm based on natural neighbor. Knowledge-Based Systems 92. https://doi.org/10.1016/j.knosys.2015.10.014.

Kalaycı, İ., Ercan, T., 2018. Anomaly detection in wireless sensor networks data by using histogram based outlier score method. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–6.

Karypis, G., Han, E.-H., Kumar, V., 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer 32 (8), 68–75. https://doi.org/10.1109/2.781637.

Kaufman, L., Rousseeuw, P., 2009. Finding Groups in Data: An Introduction to Cluster Analysis.

Keller, F., Müller, E., Böhm, K., 2012. HiCS: high contrast subspaces for density-based outlier ranking. In: 2012 IEEE 28th International Conference on Data Engineering, pp. 1037–1048.

Kirner, E., Schubert, E., Zimek, A., 2017. Good and bad neighborhood approximations for outlier detection ensembles. In: Beecks, C., Borutta, F., Kröger, P., Seidl, T. (Eds.), Similarity Search and Applications. Springer International Publishing, pp. 173–187.

Knorr, E., Ng, R., Tucakov, V., 2000. Distance-based outliers: algorithms and applications. The VLDB Journal 8, 237–253.

Knorr, E.M., Ng, R.T., 1998. Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24rd International Conference on Very Large Data Bases, pp. 392–403.

Kontaki, M., Gounaris, A., Papadopoulos, A.N., Tsichlas, K., Manolopoulos, Y., 2011. Continuous monitoring of distance-based outliers over data streams. In: 2011 IEEE 27th International Conference on Data Engineering, pp. 135–146.

Kriegel, H., Kröger, P., Zimek, A., 2009a. Outlier detection techniques. In: Proc. Tutorial KDD, pp. 1–10.

Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A., 2009b. LoOP: local outlier probabilities. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1649–1652.

Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A., 2009c. Outlier detection in axis-parallel subspaces of high dimensional data. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 831–838.

Kriegel, H.-P., Kroger, P., Schubert, E., Zimek, A., 2011. Interpreting and unifying outlier scores. In: Proceedings of the 2011 SIAM International Conference on Data Mining (SDM), pp. 13–24.

Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I., Kim, K.J., 2019. A survey of deep learning-based network anomaly detection. Cluster Computing 22 (1), 949–961. https://doi.org/10.1007/s10586-017-1117-8.

Lai, K.-H., Zha, D., Wang, G., Xu, J., Zhao, Y., Kumar, D., Chen, Y., Zumkhawaka, P., Wan, M., Martinez, D., Hu, X., 2020. TODS: an automated time series outlier detection system. CoRR, arXiv:2009.09822.

Lai, K.H., Zha, D., Xu, J., Zhao, Y., Wang, G., Hu, X., 2021a. Revisiting time series outlier detection: definitions and benchmarks. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.

Latecki, L.J., Lazarevic, A., Pokrajac, D., 2007. In: Perner, P. (Ed.), Outlier Detection with Kernel Density Functions BT - Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, pp. 61–75.

Lazarevic, A., Kumar, V., 2005. Feature bagging for outlier detection. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 157–166.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., Hu, X., 2020. COPOD: copula-based outlier detection. In: IEEE International Conference on Data Mining (ICDM).

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422.

Lozano, E., Acuna, E., 2005. Parallel Algorithms for Distance-Based and Density-Based Outliers, pp. 729–732.

Ma, X., Wu, J., Xue, S., Yang, J., Sheng, Q.Z., Xiong, H., 2021. A comprehensive survey on graph anomaly detection with deep learning. CoRR, arXiv:2106.07178.

MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations.

Micenková, B., McWilliams, B., Assent, I., 2015. Learning representations for outlier detection on a budget. CoRR, arXiv:1507.08104.

Momtaz, R., Mohssen, N., Gowayyed, M.A., 2013. DWOF: a robust density-based outlier detection approach. In: Sanches, J.M., Micó, L., Cardoso, J.S. (Eds.), Pattern Recognition and Image Analysis. Springer Berlin Heidelberg, pp. 517–525.

Moonesinghe, H.D.K., Tan, P.-N., 2008. Outrank: a graph-based outlier detection framework using random walk. International Journal on Artificial Intelligence Tools 17 (01), 19–36. https://doi.org/10.1142/S0218213008003753.

Morady, H., Lumpur, K., Suhaimi, M., Hosseinkhani, J., 2013. Outlier Detection in Stream Data by Clustering Method.

Moshtaghi, M., Bezdek, J.C., Havens, T.C., Leckie, C., Karunasekera, S., Rajasegarar, S., Palaniswami, M., 2014. Streaming analysis in wireless sensor networks. Wireless Communications and Mobile Computing 14 (9), 905–921. https://doi.org/10.1002/wcm.2248.

Moshtaghi, M., Bezdek, J.C., Leckie, C., Karunasekera, S., Palaniswami, M., 2015. Evolving fuzzy rules for anomaly detection in data streams. IEEE Transactions on Fuzzy Systems 23 (3), 688–700. https://doi.org/10.1109/TFUZZ.2014.2322385.

Muller, E., Assent, I., Steinhausen, U., Seidl, T., 2008. OutRank: ranking outliers in high dimensional data. In: 2008 IEEE 24th International Conference on Data Engineering Workshop, pp. 600–603.

Müller, E., Günnemann, S., Färber, I., Seidl, T., 2010. Discovering multiple clustering solutions: grouping objects in different views of the data. In: Proceedings - International Conference on Data Engineering, p. 1220.

Müller, E., Schiffer, M., Seidl, T., 2011. Statistical selection of relevant subspace projections for outlier ranking. In: 2011 IEEE 27th International Conference on Data Engineering, pp. 434–445.

Ng, R.T., Han, J., 1994. Efficient and effective clustering methods for spatial data mining. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 144–155.
Nguyen, H.V., Ang, H.H., Gopalkrishnan, V., 2010. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In: Kitagawa, H., Ishikawa, Y., Li, Q., Watanabe, C. (Eds.), Database Systems for Advanced Applications. Springer Berlin Heidelberg, pp. 368–383.
Ning, J., Chen, L., Chen, J., 2018. Relative density-based outlier detection algorithm. In: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, pp. 227–231.
Papadimitriou, S., Kitagawa, H., Gibbons, P.B., Faloutsos, C., 2003. LOCI: fast outlier detection using the local correlation integral. In: Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405), pp. 315–326.
Park, C.M., Jeon, J., 2015. Regression-based outlier detection of sensor measurements using independent variable synthesis. In: Proceedings of the Second International Conference on Data Science - Volume 9208, pp. 78–86.
Pasillas-Díaz, J., Ratté, S., 2016. Bagged subspaces for unsupervised outlier detection: FBSO. Computational Intelligence 33. https://doi.org/10.1111/coin.12097.
Pincus, R., 1995. Barnett, V., and Lewis T.: Outliers in Statistical Data. 3rd edition. J. Wiley & Sons 1994, XVII. 582 pp., £49.95. Biometrical Journal 37 (2), 256. https://doi.org/10.1002/bimj.4710370219.
Pokrajac, D., Lazarevic, A., Latecki, L.J., 2007. Incremental local outlier detection for data streams. In: 2007 IEEE Symposium on Computational Intelligence and Data Mining, pp. 504–515.
Porwal, U., Mukund, S., 2018. Credit card fraud detection in e-commerce: an outlier detection approach. arXiv:1811.02196. [Online]. Available: https://arxiv.org/abs/1811.02196.
Qin, X., Cao, L., Rundensteiner, E.A., Madden, S., 2019. Scalable Kernel Density Estimation-Based Local Outlier Detection over Large Data Streams. EDBT.
Radovanović, M., Nanopoulos, A., Ivanović, M., 2015. Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Transactions on Knowledge and Data Engineering 27 (5), 1369–1382. https://doi.org/10.1109/TKDE.2014.2365790.
Ramakrishnan, J., Shaabani, E., Li, C., Sustik, M.A., 2019. Anomaly detection for an E-commerce pricing system. CoRR, arXiv:1902.09566.
Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. SIGMOD Record 29 (2), 427–438. https://doi.org/10.1145/335191.335437.
Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., Samatova, N.F., 2015. Anomaly detection in dynamic networks: a survey. WIREs: Computational Statistics 7 (3), 223–247. https://doi.org/10.1002/wics.1347.
Rayana, S., Akoglu, L., 2016. Less is more: building selective anomaly ensembles. ACM Transactions on Knowledge Discovery from Data 10 (4). https://doi.org/10.1145/2890508.
Ren, D., Rahal, I., Perrizo, W., Scott, K., 2004a. A vertical distance-based outlier detection method with local pruning. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 279–284.

Ren, D., Wang, B., Perrizo, W., 2004b. RDF: a density-based outlier detection method using vertical data representation. In: Fourth IEEE International Conference on Data Mining (ICDM'04), pp. 503–506.
Rizk, H., Elgokhy, S., Sarhan, A., 2015. A hybrid outlier detection algorithm based on partitioning clustering and density measures. In: 2015 Tenth International Conference on Computer Engineering Systems (ICCES), pp. 175–181.
Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M., 2018. Deep one-class classification. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, vol. 80. PMLR, pp. 4393–4402. http://proceedings.mlr.press/v80/ruff18a.html.
Saha, B.N., Ray, N., Zhang, H., 2009. Snake validation: a PCA-based outlier detection method. IEEE Signal Processing Letters 16 (6), 549–552. https://doi.org/10.1109/LSP.2009.2017477.
Salehi, M., Leckie, C.A., Moshtaghi, M., Vaithianathan, T., 2014. A relevance weighted ensemble model for anomaly detection in switching data streams. In: Tseng, V.S., Ho, T.B., Zhou, Z.-H., Chen, A.L.P., Kao, H.-Y. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer International Publishing, pp. 461–473.
Samparthi, V., Verma, H., 2010. Outlier detection of data in wireless sensor networks using kernel density estimation. International Journal of Computer Applications 5. https://doi.org/10.5120/924-1302.
Satman, M.H., 2013. A new algorithm for detecting outliers in linear regression. International Journal of Statistics and Probability 2, 101.
Schubert, E., Wojdanowski, R., Zimek, A., Kriegel, H.-P., 2012. On evaluation of outlier rankings and outlier scores. In: Proceedings of the 2012 SIAM International Conference on Data Mining (SDM), pp. 1047–1058.
Schubert, E., Zimek, A., Kriegel, H.-P., 2014. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery 28 (1), 190–237. https://doi.org/10.1007/s10618-012-0300-z.
Sehwag, V., Chiang, M., Mittal, P., 2021. SSD: a unified framework for self-supervised outlier detection. CoRR, arXiv:2103.12051.
Shahid, N., Naqvi, I.H., Qaisar, S.B., 2015. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey. Artificial Intelligence Review 43, 193–228. https://doi.org/10.1007/s10462-012-9370-y.
Sheikholeslami, G., Chatterjee, S., Zhang, A., 2000. WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal 8 (3), 289–304. https://doi.org/10.1007/s007780050009.
Shekhar, S., Shah, N., Akoglu, L., 2020. FAIROD: fairness-aware outlier detection. CoRR, arXiv:2012.03063.
Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: a data mining perspective. SIGKDD Explorations Newsletter 19 (1), 22–36. https://doi.org/10.1145/3137597.3137600.
Shukla, M., Kosta, Y.P., Chauhan, P., 2015. Analysis and evaluation of outlier detection algorithms in data streams. In: 2015 International Conference on Computer, Communication and Control (IC4), pp. 1–8.

Singh, G., Masseglia, F., Fiot, C., Marascu, A., Poncelet, P., 2010. Mining common outliers for intrusion detection. In: Guillet, F., Ritschard, G., Zighed, D.A., Briand, H. (Eds.), Advances in Knowledge Discovery and Management. In: Studies in Computational Intelligence, vol. 292. Springer, Berlin, Heidelberg.

Smrithy, G.S., Munirathinam, S., Balakrishnan, R., 2016. Online anomaly detection using non-parametric technique for big data streams in cloud collaborative environment. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1950–1955.

Su, S., Xiao, L., Ruan, L., Gu, F., Li, S., Wang, Z., Xu, R., 2019. An efficient density-based local outlier detection approach for scattered data. IEEE Access 7, 1006–1020. https://doi.org/10.1109/ACCESS.2018.2886197.

Tamboli, J., Shukla, M., 2016. A survey of outlier detection algorithms for data streams. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3535–3540.

Tang, B., He, H., 2017. A local density-based approach for outlier detection. Neurocomputing 241, 171–180. https://doi.org/10.1016/j.neucom.2017.02.039.

Tang, J., Chen, Z., Fu, A.W., Cheung, D.W., 2002. Enhancing effectiveness of outlier detections for low density patterns. In: Chen, M.-S., Yu, P.S., Liu, B. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, pp. 535–548.

Tang, X., Yuan, R., Chen, J., 2015. Outlier detection in energy disaggregation using subspace learning and Gaussian mixture model TT. International Journal of Control and Automation 8 (8), 161–170. https://www.earticle.net/Article/A253913.

Ting, Ming, Kai, Xu, Bi-Cun, Washio, Takashi, Zhou, Zhi-Hua, 2020. Isolation distributional kernel: a new tool for kernel based anomaly detection. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 198–206.

Tran, Luan, Fan, Liyue, Shahabi, Cyrus, 2016. Distance-based outlier detection in data streams. Proceedings of the VLDB Endowment 9 (12), 1089–1100. https://doi.org/10.14778/2994509.2994526.

Tukey, John, 2020. Exploratory Data Analysis. Pearson Modern Classics.

Uddin, M.S., Kuh, A., Weng, Y., Ilić, M., 2015. Online bad data detection using kernel density estimation. In: IEEE Power Energy Society General Meeting, pp. 1–5.

van Hieu, D., Meesad, P., 2016. A fast outlier detection algorithm for big datasets. In: Meesad, P., Boonkrong, S., Unger, H. (Eds.), Recent Advances in Information and Communication Technology. Springer International Publishing, pp. 159–169.

van Stein, B., van Leeuwen, M., Bäck, T., 2016. Local subspace-based outlier detection using global neighbourhoods. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1136–1142.

Verma, S., Ernst, M., Just, R., 2021. Removing biased data to improve fairness and accuracy. ArXiv. arXiv:2102.03054 [abs].

Vu, N.H., Gopalkrishnan, V., 2009. Efficient pruning schemes for distance-based outlier detection. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, pp. 160–175.

Walfish, S., 2006. A review of statistical outlier methods. Pharmaceutical Technology 30.

Wang, C., Gao, H., Liu, Z., Fu, Y., 2018a. A new outlier detection model using random walk on local information graph. IEEE Access 6, 75531–75544. https://doi.org/10.1109/ACCESS.2018.2883681.

Wang, W., Yang, J., Muntz, R., 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining. VLDB.

Wang, X., Wang, X.L., Ma, Y., Wilkes, D.M., 2015. A fast MST-inspired KNN-based outlier detection method. Information Systems 48, 89–112. https://doi.org/10.1016/j.is.2014.09.002.

Wu, K., Zhang, K., Fan, W., Edwards, A., Yu, P.S., 2014. RS-forest: a rapid density estimator for streaming anomaly detection. In: 2014 IEEE International Conference on Data Mining, pp. 600–609.

Xiao, T., Zhang, C., Zha, H., 2015. Learning to detect anomalies in surveillance video. IEEE Signal Processing Letters 22 (9), 1477–1481. https://doi.org/10.1109/LSP.2015.2410031.

Yang, D., Rundensteiner, E.A., Ward, M.O., 2009a. Neighbor-based pattern detection for windows over streaming data. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 529–540.

Yang, X., Latecki, L.J., Pokrajac, D., 2009b. Outlier detection with globally optimal exemplar-based GMM. In: Proceedings of the 2009 SIAM International Conference on Data Mining (SDM), pp. 145–154.

Yilmaz, S.F., Kozat, S.S., 2020. PySAD: a streaming anomaly detection framework in python. CoRR, arXiv:2009.02572.

Yin, J., Wang, J., 2016. A model-based approach for text clustering with outlier detection. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pp. 625–636.

Yiyong, J., Jifu, Z., Jianghui, C., Sulan, Z., Lihua, H., 2007. The outliers mining algorithm based on constrained concept lattice. In: The First International Symposium on Data, Privacy, and E-Commerce (ISDPE 2007), pp. 80–85.

Zahn, C.T., 1971. Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers C-20 (1), 68–86. https://doi.org/10.1109/T-C.1971.223083.

Zha, D., Lai, K.-H., Wan, M., Hu, X., 2020. Meta-AAD: active anomaly detection with deep reinforcement learning. CoRR, arXiv:2009.07415.

Zhang, Haibin, Liu, Jiajia, Zhao, Cheng, 2016. Distance Based Method for Outlier Detection of Body Sensor Networks, WS, EAI.

Zhang, J., 2013. Advancements of outlier detection: a survey. ICST Transactions on Scalable Information Systems 13, e2. https://doi.org/10.4108/trans.sis.2013.01-03.e2.

Zhang, J., Hsu, W., Li Lee, M., 2005. Clustering in dynamic spatial databases. Journal of Intelligent Information Systems 24 (1), 5–27. https://doi.org/10.1007/s10844-005-0265-0.

Zhang, J., Jiang, Y., Chang, K.H., Zhang, S., Cai, J., Hu, L., 2009a. A concept lattice based outlier mining method in low-dimensional subspaces. Pattern Recognition Letters 30 (15), 1434–1439. https://doi.org/10.1016/j.patrec.2009.07.016.

Zhang, K., Hutter, M., Jin, H., 2009b. A new local distance-based outlier detection approach for scattered real-world data. CoRR, arXiv:0903.3257.
Zhang, L., Lin, J., Karim, R., 2018. Adaptive kernel density-based anomaly detection for nonlinear systems. Knowledge-Based Systems 139, 50–63. https://doi.org/10.1016/j.knosys.2017.10.009.
Zhao, Y., Hryniewicki, M.K., 2019a. DCSO: dynamic combination of detector scores for outlier ensembles. CoRR, arXiv:1911.10418.
Zhao, Y., Hryniewicki, M.K., 2019b. XGBOD: improving supervised outlier detection with unsupervised representation learning. CoRR, arXiv:1912.00290.
Zhao, Y., Hryniewicki, M.K., Nasrullah, Z., Li, Z., 2018. LSCP: locally selective combination in parallel outlier ensembles. CoRR, arXiv:1812.01528.
Zhao, Y., Hu, X., Cheng, C., Wang, C., Xiao, C., Wang, Y., Sun, J., Akoglu, L., 2020. SUOD: a scalable unsupervised outlier detection framework. CoRR, arXiv:2003.05731.
Zheng, Z., Jeong, H., Huang, T., Shu, J., 2016. KDE based outlier detection on distributed data streams in multimedia network. Multimedia Tools and Applications 76, 18027–18045.
Zhou, C., Paffenroth, R.C., 2017. Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 665–674.
Zimek, A., Campello, R.J.G.B., Sander, J., 2014a. Data perturbation for outlier detection ensembles. In: Proceedings of the 26th International Conference on Scientific and Statistical Database Management.
Zimek, A., Campello, R.J.G.B., Sander, J., 2014b. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. SIGKDD Explorations Newsletter 15 (1), 11–22. https://doi.org/10.1145/2594473.2594476.
Zimek, A., Gaudet, M., Campello, R.J.G.B., Sander, J., 2013. Subsampling for efficient and effective unsupervised outlier detection ensembles. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 428–436.

This page intentionally left blank

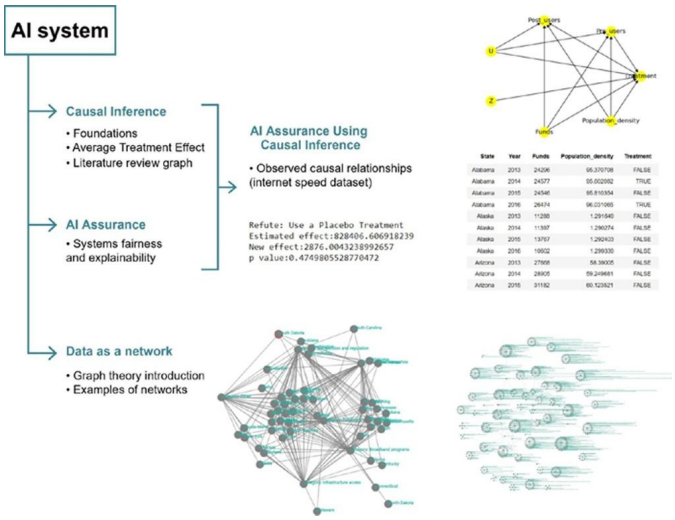
8

AI assurance using causal inference: application to public policy

Andrei Svetovidov[a], Abdul Rahman[b], and Feras A. Batarseh[c]

[a]*Commonwealth Cyber Initiative, Virginia Tech, Arlington, VA, United States* [b]*Deloitte & Touche, LLP, Baltimore, MD, United States* [c]*Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States*

Graphical abstract



Abstract

Developing and implementing AI-based solutions help state and federal government agencies, research institutions, and commercial companies enhance decision-making processes, automate chain operations, and reduce the consumption of natural and human resources. At the same time, most AI approaches used in practice can only be represented as "black boxes" and suffer from the lack of transparency. This can eventually lead to unexpected outcomes

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00021-4

and undermine trust in such systems. Therefore it is crucial not only to develop effective and robust AI systems, but to make sure their internal processes are explainable and fair. Our goal in this chapter is to introduce the topic of designing assurance methods for AI systems with high-impact decisions using the example of the technology sector of the US economy. We explain how these fields would benefit from revealing cause-effect relationships between key metrics in the dataset by providing the causal experiment on technology economics dataset. Several causal inference approaches and AI assurance techniques are reviewed and the transformation of the data into a graph-structured dataset is demonstrated.

Keywords

Causality, network data, public policy, AI assurance

Highlights

- Fundamentals of the causal inference theory
- Review of the concept of AI assurance and its connection to causality
- Assurance-focused causal experiment on the internet speed dataset
- Methods of graph-based data representation and analysis

8.1 Introduction and motivation

Artificial Intelligence (AI) has experienced a tremendous growth trend during the last decade due to availability of multivariate large-scale datasets and advancements in high-performance computations with multi-core GPUs. AI methods, such as classic machine learning algorithms, deep neural networks, and reinforcement learning, demonstrated impressive results in solving prediction and classification tasks in many domains, including transportation, healthcare, and finance (Boire, 2018). However, utilization of such methods is heavily underexplored in policy making. The institutions and agencies participating in legislature activities would clearly benefit from using cutting-edge AI models to make the lawmaking process more effective and better able to address goals of government and state level officials (Zuidewijk et al., 2021).

At the same time, the lawmaking process itself is complex in nature, involves multiple steps, and can also vary from state to state or even at higher granularity. Moreover, the decisions taken by officials in the form of issued policies determine the path of development of the country and its inhabitants. We have seen many examples of how passed laws influenced the lives of millions of people, such as the Coronavirus Act (Coronavirus Preparedness and Response Supplemental Appropriations Act, 2020). The law facilitated the production and distribution of COVID-19 tests and vaccines and allocated $22 billion in funding, which led to a faster vaccination across the United States, and therefore building collective immunity against the disease. Since AI-based legislation systems are to be built on such sensitive and influential information, it is very important to assure their trustworthiness, transparency, and fairness towards the residents. If it is possible to explain how AI methods work and why they generate such results, we could maintain public trust in them and would eventually fully integrate such systems into policy making cycles.

One of the ways to leverage AI assurance in this scenario is to consider causal inference methods applied to the metrics of interest. For instance, can we infer that a cause-effect relationship exists between the proposed COVID-19 vaccine distribution law and the number of current positive cases? Another example is whether the regulations issued by the U.S. Department of Transportation lead to reduced driving times on target roads. Such dependencies between different factors are typically not captured by classic machine learning algorithms, and more sophisticated methods are required. In addition, it would be helpful to also consider dependencies between input data vectors based on some other contextual information. For example, the laws approved and passed in separate states might influence each other to some degree if the states are geographically close to each other, situated in the same region of the U.S., or governed by the same political force. The aspects mentioned above should be carefully considered while creating an AI assurance model for policy making (Perry and Uuk, 2019).

In the next section of this chapter, we introduce the concept of causal inference and provide a detailed overview of some of the methods with an accompanying knowledge graph.

8.2 Causal inference

Machine Learning (ML) and Deep Learning (DL) algorithms are heavily used as core elements of AI systems since they bring a power of detecting input/output data relationships and predicting the outcome for future inputs. However, a significant limitation of ML models is their inability to account for mutual correlation or association between inputs and outputs, leaving aside more complex dependencies. Thus cause-effect relationships cannot be captured by a feed-forward neural network, and applying this approach directly to some problems can lead to false inferences. Let us take one example: a user notices that their screen freezes every time they open Google Chrome on a personal computer. The user may assume that the problem is caused by launching the browser, whereas the underlying reason might be that there are too many active background OS processes that utilize most of available RAM, so that the system cannot easily handle such a "RAM-hungry" web browser as Chrome. If our neural network is trained on the results of user experience surveys to predict whether the system would freeze, it would produce wrong predictions for users whose RAM is not overloaded with running OS processes.

The example provided above clearly proves the famous statement: "*correlation does not imply causation*" (Aldrich, 1995). Indeed, there is an obvious correlation between crashing an OS and launching a web browser, although one does not cause another. In other settings, such as healthcare or national security, such incorrect inferences may lead to dramatic consequences involving safety and lives of people, as mentioned by Hamid (2016), so it is necessary to understand causal relationships before applying common AI solutions. In Section 8.2.1, we introduce the concepts of causality and causal inference and their mathematical foundations.

8.2.1 An introduction to causal inference

Causal inference is a relatively new field of study within AI context, but showed to be very promising in recent years. Although first inference engines were introduced as a main block of expert systems in 1970–1980s (Hayes-Roth et al., 1983), they were not truly intelligent; they relied on predefined logical rules and were not capable of performing data-driven knowledge inference.

Let us start with explaining the difference in terms used in this area. *Causality* is an existing relationship between the effect and what it was caused by (object, state, or process). *Causation* means the act of causing a particular event/state/process (Honderich, 1988). Although these two terms are often used interchangeably, *causation* can be viewed as a process of initiation of *causality*. For instance, the rain caused city dwellers to take out and open their umbrellas, which shows the relationship of causality between these events. The causation occurred right after the rain started, making people use their umbrellas. *Causal Inference* is a field of study that attempts to reveal causal relationships between nodes via making causal assumptions (Pearl, 2009).

Causal inference methods were heavily involved in addressing challenges in healthcare (Moser et al., 2020); therefore they inherited some terminology from the latter. In addition to standard machine learning concepts of input data X, known as *covariate features of a patient*, and output data Y, known as *outcome*, we also introduce T—*treatment*—the action taken on a patient (or, in medical terms, a treatment that was given to them). The relationships between X, T, and Y are usually presented as a directed acyclic graph (DAG). In general, DAG represents complex causal structures, can be very cumbersome, and includes multiple nodes of each type, but for the sake of simplicity, let us consider the following scenario: how a certain medication affects the patient's blood pressure level (Szolovits and Sontag, 2019). Here, X is the information about the patient known beforehand, T is a medication, which can be either $T0$ (control treatment) or $T1$ (actual medication), and Y is the patient's blood pressure after being treated. The corresponding causal graph is shown in Fig. 8.1.

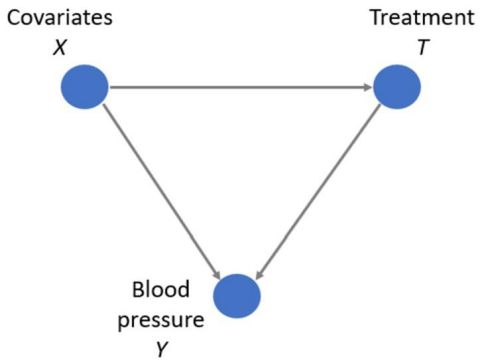


FIGURE 8.1 Causal DAG example.

For each individual, we want to understand if the hypertension treatment actually works. For this we need to compare the effects of each treatment on a patient, which can be done via calculating conditional average treatment effect as per Neyman-Rubin causal model (Sekhon, 2008):

$$CATE\,(x_i) = E_{Y_1 \sim p(Y_1|x_i)}\,[x_i] - E_{Y_0 \sim p(Y_0|x_i)}\,[x_i]\,, \tag{8.1}$$

where $E_{Y_1 \sim p(Y_1|x_i)}\,[x_i]$, and $E_{Y_0 \sim p(Y_0|x_i)}\,[x_i]$ is the expectation of the outcome had the individual x_i had and had not been treated, respectively. x_i is a set of features of the patient.

The average treatment effect for the entire population can be calculated as the expectation over CATE values for all instances:

$$ATE = E_{x \sim p(x)}[CATE(x_i)] \tag{8.2}$$

However, it is often impossible to measure the outcome of both treatments applied to the same individual due to many reasons, including safety and ethics, which is known as the fundamental problem of causal inference (Holland, 1986). Hospitals make educated decisions on what treatment to deliver to each patient, and providing several treatments simultaneously can not only ruin the reputation of the doctor or hospital and healthcare system in general, but lead to malicious effects on patient's health and violate certain regulations. Therefore causal inference based on counterfactuals is only dealing with the data obtained from control and treatment

groups (Szolovits and Sontag, 2019):

$$ATE = E_{x \sim p(x)}\left[E[Y_1 \mid x, T = 1] - E[Y_0 \mid x, T = 0]\right], \tag{8.3}$$

where Y_1 and Y_0 are the responses of patients being provided T_1 and T_0, respectively.

As explained by Szolovits and Sontag (2019), this approach requires several strong assumptions to be held:

- Ignorability: there are no unobserved confounding variables. The mathematical form of this assumption shows that the outcome is independent of treatment given input data:

$$(Y_0, Y_1) \perp T \mid x \tag{8.4}$$

- Common support: there is always a stochasticity in treatment decisions:

$$p(T = t \mid X = x) > 0 \; \forall t, x \tag{8.5}$$

  For example, we have a subpopulation of individuals with red hair. These should include patients from both control and treatment groups to satisfy this assumption.
- Stable unit treatment value assumption (SUTVA): the response (outcome) for a particular individual to a provided treatment is independent of treatments of other units.

Once the conditions mentioned above are satisfied, we can attempt to apply ML in a standard way to draw the relationships between inputs and outputs. It should be mentioned that these assumptions do not perfectly hold in the real world, and inferring CATE values is still a probabilistic process. In Section 8.2.2, we overview state-of-the-art causal inference approaches that address these limitations.

8.2.2 Overview of causal inference methods

In Section 8.2.1, we described the conditions that need to hold to apply ML methods for causal inference. One of the first approaches for estimating average treatment effect (ATE) is called covariance adjustment (Szolovits

and Sontag, 2019). In this method, a parametric model is being fitted on training data:

$$f(x,t) \approx E[Y_t \mid T = t, x], \tag{8.6}$$

where $f(x,t)$ is the function for approximation of the expectation $E[Y_t \mid T = t, x]$, and t equals 0 or 1 in binary setting.

Once the model is trained, we can calculate ATE estimation via finding the average difference between function values for each patient with treatment values 1 and 0 accordingly:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0) \tag{8.7}$$

In the simplest case, there is a linear dependency between outcome and covariates/treatment described by the following equation:

$$Y_t(x) = \alpha x + \beta t + \gamma, \tag{8.8}$$

where α, β, γ are trainable parameters.

Although this linear regression model is comparatively easy to train, it may oversimplify true causal relationships and eventually lead to incorrect assumptions. It is very important to ensure that the model has sufficient representative power to infer correct outcomes. Another reason for making incorrect assumptions—which is very common in practice—is violation of common support. In many domains and for many reasons it may not be feasible to draw data from the randomized control trial type of experiment, where the patients are assigned treatment randomly. Therefore there is an approach to estimate ATE, while mitigating the selection bias, which is called propensity score matching (Szolovits and Sontag, 2019). It requires the following steps:

1. Estimate propensity scores for each individual, i.e., define the groups of similar patients in terms of provided treatment: $p(T = t \mid x)$.
2. Match propensity scores: there are several methods, but the most popular one is based on nearest neighbor search.

3. Evaluate the matching technique, and apply a different one if the results are unsatisfactory.
4. Calculate ATE for each stratum with the formula mentioned in Section 8.2.1 (group of individuals with close propensity scores) and take average or weighted average of them to estimate ATE for the entire population.

If the assumption of *ignorability* holds, i.e., data collection being independent of missing data, this method can balance initial bias in treatment assignment prior to estimating ATE.

The precursor of causal inference methods was the developed model called Bayesian network. It represented conditional dependencies of variables in a form of DAG. Even though Bayesian DAG cannot be used directly to identify causal effect, it gave rise to further developments in causality. One of relatively early causal inference methods, Bayesian additive regression trees (BART), uses decision tree algorithm as a building block, where the path is determined by conditions on X and T, and the value of Y is found at the end point of each path (Hill, 2011). Aggregate outcome from separate single trees is deemed a final result. This method predicted ATE more accurately than linear regression and propensity score matching, but was very sensitive to the amount of training data. Künzel et al. (2019) proposed a more sophisticated method, in which the model consists of base learners: BART and causal forests. This approach provides a richer representation, thanks to estimating an outcome for a treated individual using control-outcome estimator, and vice versa, which helps to account for imbalance between treatment groups.

Better results can be achieved with deep learning. Johansson et al. (2016) applied deep neural networks to counterfactual inference. Authors presented a modified version of this approach, called TARNet in Shalit et al. (2017), where the network was augmented to the neural network with two parallel blocks to estimate the effect under treatment and control, respectively. Also, the treatment assignment bias is adjusted by adding IPM (integral probability metric) term, and the final objective function is a trade-off between treatment imbalance and accuracy. The technique proposed by Shalit et al. (2017) was utilized in other works, including (Schwab et

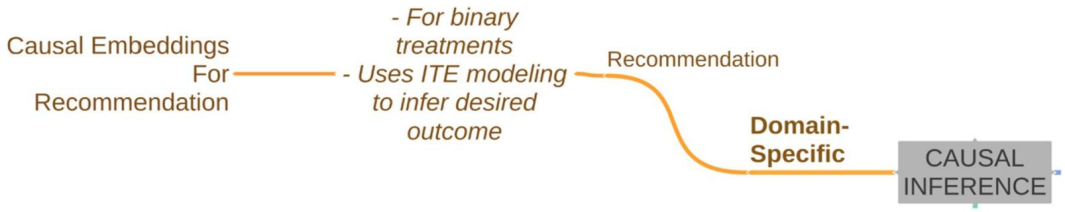


FIGURE 8.2 Domain-specific part of literature review knowledge graph.

al., 2018), where TARNet was extended to the algorithm with the ability to handle multiple different (non-binary) treatments and augment insufficient input data. Schwab et al. (2020) improved this work, where the joint neural network considers "dosage," or real-value treatment. The network structure of the dataset was considered in Guo et al. (2020) where TARNet was combined with GNN to predict ATE. Yao et al. (2018) also took into account connections between instances (individuals) via implementation of local similarity information along with treatment distribution balancing and deep learning ATE estimator.

In multiple works, researchers extended ideas of counterfactual inference to more advanced cases. Hartford et al. (2017) introduced a two-stage DNN-based model, which accounts for presence of hidden confounders via using instrumental variables. Lim et al. (2018) developed a framework for predicting an outcome of a chain of treatments. Kobrosly (2020) developed a Python package to estimate a dose-response curve with the generalized propensity score and targeted maximum likelihood estimation. Louizos et al. (2017) built a causal effect variational autoencoder to combine advancements of latent variables in machine learning and proxy variables utilization in causality. Rakesh et al. (2018) extended this work to also consider the pairwise spillover effect between covariates. Some of the works represent domain adaptations of causal inference methods, such as Bonner and Vasile (2018), where recommendation policy optimization is done via increasing the desired outcome with ITE modeling.

The graphical representation of this literature overview is demonstrated in Fig. 8.2 and Fig. 8.3. All the works are placed into groups in accordance with their relation to causal inference:

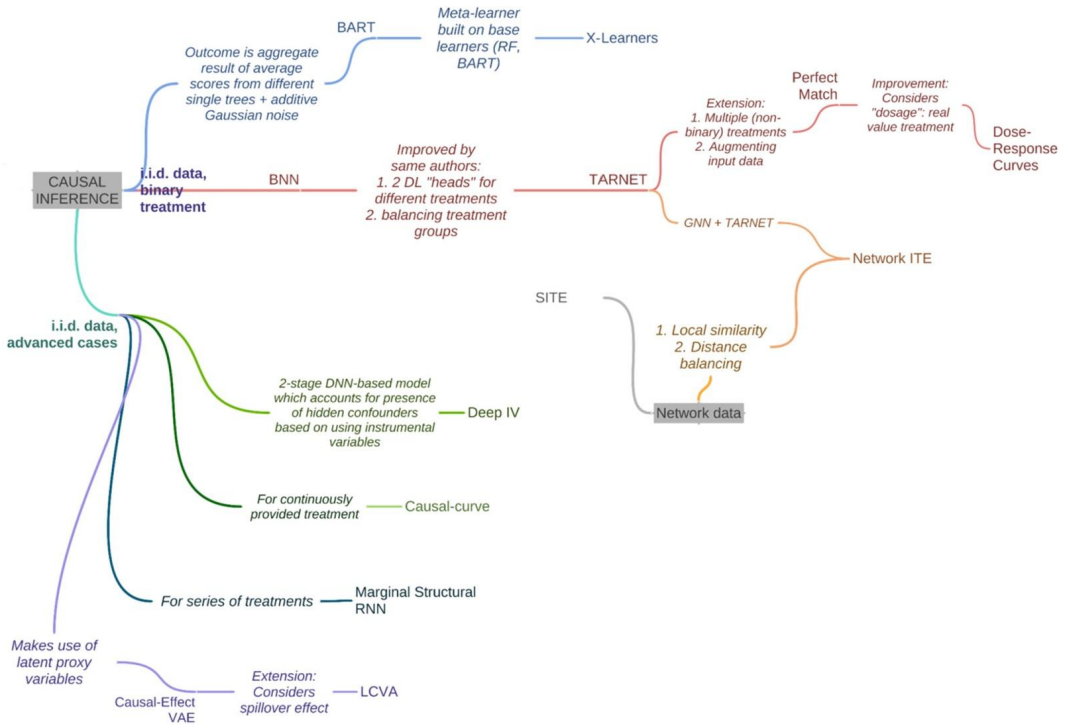


FIGURE 8.3 Literature review knowledge graph (except domain-specific part).

- Independent and identically distributed (i.i.d.) data with binary treatment
- i.i.d. data with non-binary treatment (advanced cases)
- Domain-specific

The graph reveals a "parent-child" dependency of papers based on implementation and modification of ideas of some works in others. Edges of the connecting lines are also accompanied with short work description (in non-bold italic). For the sake of readability, the graph is broken down into two parts.

8.3 AI assurance using causal inference

In the previous section, we presented the fundamentals of causal inference and introduced a review of some recent works in the area. This section fa-

miliarizes the reader with the main ideas of AI assurance (Section 8.3.1), and the authors attempt to connect these two areas of study by providing an experiment on tech-econ policy dataset (Section 8.3.2).

8.3.1 AI assurance: goals and methods

As we mentioned earlier, assuring AI systems is becoming more necessary due to the growing demand in such systems, as well as increasing requirements to their transparency. For many critical applications, such as healthcare or military operations, there are several key assurance pillars that AI engineers should consider during building, training, and testing algorithms (Batarseh et al., 2021).

First, the outcomes produced by an AI system should be trustworthy, otherwise the fundamental idea of replacing certain human-centered operations with automated intelligent systems would be undermined, which would prevent their further development. Second, the decisions made by the system should be fair and ethical with regards to its users, or to be able to detect, avoid, or eliminate any sort of bias in its outputs. We also want the AI system and its processes to be explainable to in turn ensure trustworthiness, but also be secured from potential outside threats. In certain scenarios, some assurance goals are more important than others. For instance, in military applications ethics gives way to safety, but in civil applications, such as automated hiring, fairness comes first.

Since the main components in the AI pipeline are the model itself and input data/outcomes, they should be the main target of assurance methods. Input data should be checked for completeness and importance for a particular task the AI system is created for. Other approaches aim at revealing details of internal processes in the algorithm during the training phase. For instance, deep neural network training mechanism is based on updating its connection weights, and controlling the changes of weight values is one of the ways of providing explainability. Moreover, the researchers can develop assurance metrics for each of the categories, which can be integrated directly into the objective function. This method largely depends on what AI approach is used and how the model is trained.

In the following section we apply the basics of causal inference to the assurance problem.

8.3.2 Methods for leveraging causality in assurance

In Section 8.2, we presented some causal inference methods. Although the goal of most of these studies is to perform an accurate counterfactual inference, where developed algorithms are used for prediction, we can look at causal inference for assurance purposes from a different angle.

As we mentioned earlier, there are 3 strong assumptions that need to be made to find ATE based on counterfactuals. The common support property can be utilized by itself to address the questions of fairness and ethics. This property states that there should be no bias in treatment decisions, and if this property holds, we can assert that the assurance issues of fairness and ethics are addressed via ensuring diversity in deciding what treatments are provided for various patients.

A more comprehensive way to conduct assurance analysis is to build a causal graph and detect connections between features. Such analysis can be useful both at the initial stage of developing an AI model and during its utilization. In the first case, a causal graph can serve as a validation tool for a less explainable AI model and make an educated decision of what features to use as inputs and outputs for this model. This method is in a way similar to imposing constraints on data/model based on real-world knowledge, such as physics-guided architecture of neural networks, as proposed by Daw et al. (2020). Section 8.3.3 describes an experiment performed to leverage this idea.

In the second scenario, the information about causal relationships in the dataset can be analyzed together with the model outputs to provide insights on the meaning of the outcomes given input data. For instance, a parallel metamodel can be run on the main model's outputs to perform calculations of metrics "on the fly" for dynamic analysis, such as mean squared errors and average output values.

In Section 8.3.3, we demonstrate a causal experiment inspired by the first approach, where the feature-focused causal model is created to define inputs and outputs for the future model.

Table 8.1 Dataset summary statistics.

Feature name	Number of non-empty records
State	400
Year	400
Funds	300
Population density	400
Internet users	250
Treatment	400

8.3.3 Application of causality in assurance: economy of technology example

The goal of the experiment is to perform causal analysis to understand how issued tech policies affect the target metric. DoWhy library (Sharma and Kiciman, 2020) was used to drive causation in a technology policy dataset (Anuga et al., 2021). The main sources of information were the Federal Communications Commission (FCC) (https://www.fcc.gov/) and U.S. Census Bureau (https://www.census.gov/) websites, which include tech-related laws passed in the U.S. on a state level in a particular year, technology metrics (such as number of internet users and average internet speeds), and characteristics of each state.

The aggregate numeric economy of technology dataset includes 341 features divided into 2 broad categories: Environmental Descriptors and Technology Metrics. The former represents contextual information about each state, such as population across different years, land/water areas, and funds available. Technology Metrics include various indicators of prevalence of internet and other technology across different age groups and devices. Another dataset includes law texts, together with state and year issued. In this study we combined and wrangled those datasets to match our needs of causal analysis. The statistics of the dataset can be found in Table 8.1.

After the dataset had been wrangled, we had the following variables:

- *State:* the name of the state
- *Year:* the year the law was passed
- *Funds:* the total amount of money spent on tech development (rendered in millions)

Table 8.2 Tech policy dataset.

	State	Year	Funds	Population_density	Treatment	Pre_users	Post_users
1	Alabama	2013	24,296	95.370708	FALSE	3.03E+06	2.96E+06
2	Alabama	2014	24,577	95.602082	TRUE	2.87E+06	3.21E+06
3	Alabama	2015	24,546	95.810354	FALSE	2.96E+06	3.46E+06
4	Alabama	2016	26,474	96.031065	TRUE	3.21E+06	3.51E+06
5	Alaska	2013	11,288	1.291649	FALSE	5.51E+05	5.21E+05
6	Alaska	2014	11,397	1.290274	FALSE	5.26E+05	5.33E+05
7	Alaska	2015	13,767	1.292403	FALSE	5.21E+05	5.38E+05
8	Alaska	2016	10,602	1.299339	FALSE	5.33E+05	5.12E+05
9	Arizona	2013	27,668	58.39005	FALSE	4.32E+06	4.10E+06
10	Arizona	2014	28,905	59.249681	FALSE	4.03E+06	4.47E+06
11	Arizona	2015	31,182	60.123521	FALSE	4.10E+06	4.91E+06

- *Population density*
- *Internet users:* the overall number of internet users of ages 3 and above
- *Treatment*: a binary label indicating whether a tech law was issued (True) or not (False)

In this case, (state, year) tuple is a unique record for our causal model. We utilize *Internet users* variable as the outcome, *Treatment* as treatment provided, and *Funds* and *Population density* as covariates. To consider the dimension of time, we also added *Pre users* and *Post users* columns, which contain the number of people using the internet in the preceding and following years accordingly. The rationale behind it is the assumption we make regarding the causal structure: the number of users in the past causes the initiation of the law (*Treatment*), which in turn causes the change in number of users in the future. To handle gaps in *Internet users* column, we implemented cubic interpolation to fill in gaps in 2013 and 2015 and padding to do so in 2017 (the value similar to the one in 2016 just means the absence of change in key metric from 2016 to 2017). The table excerpt is shown in Table 8.2, where the total number of records is 300.

Next, we initialized the causal model with the structure demonstrated in Fig. 8.4. In addition to our main nodes, we also accounted for potential unobserved confounders (U) and instrumental variables (Z). Similarly to other covariate features (*Funds, Population_ density*) U variable affects treatment

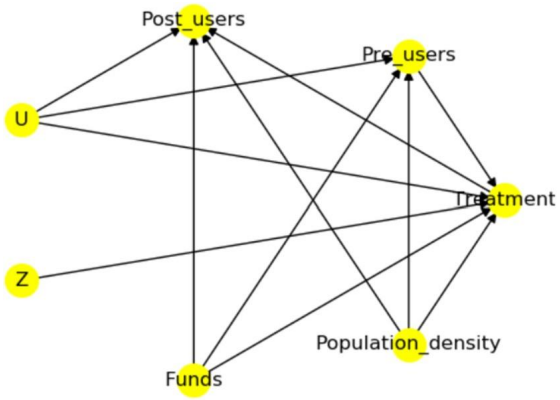


FIGURE 8.4 Causal graph.

and outcome variables: *Post_users, Pre_users*, and *Treatment*. To consider a sequence of events in time as mentioned earlier, we modeled the influence of the number of users in the past (*Pre_users*) on *Treatment*, which in turn affects *Post_users*. The Z variable directly affects only *Treatment* in this experiment.

Then we identified the causal effect qualitatively based on provided dependencies, where the assumption of little importance of unobserved confounders was made. The produced estimate has been utilized to obtain the estimated average treatment effect. We chose the propensity score matching method. DoWhy also supports the "refute estimate" method, allowing researchers to compare the results with the placebo treatment. In this experiment, treatment values (TRUE/FALSE) were replaced with random binary values to emulate causal effect after treatment random permutation. The output is as follows:

```
Refute: Use a Placebo Treatment
Estimated Effect: 828406.606918239
New Effect: 2876.0043238992657
p value: 0.4749805528770472
```

As we can see from the numbers, the release of a tech policy leads to an increase of the number of internet users by 828,406 on average compared to 2876 in the case of placebo treatment (can be viewed as "idle" legislation).

Although the number is not an exact indicator of quantitative effect and varies from one legislation to another, the model is consistent. Based on the results, we can conclude that the number of internet users is a good choice of an output for an AI learning algorithm, whereas other variables can be utilized as inputs given time dimension constraints.

8.4 Network representations of data

For many applications it can be beneficial to see the dataset as a network structure through revealing connections between entities. Preserving internal dependencies is one example of how to facilitate in-depth analysis of the dataset and ensure higher transparency during data-related AI stages. We provide an overview of the basic provisions of graph theory and recurrent graph neural networks (RGNN), and finally provide several graph representations of U.S. states and corresponding technology laws from the dataset used in Section 8.3.

8.4.1 An introduction to graph theory

In the following section, essential concepts of graph theory are embodied within the concepts of graph neural networks (GNN), which are crucial for building and training a ML algorithm that handles graph data most effectively (Scarselli et al., 2009; Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017). The core input data structure for a GNN to work is the graph. As alluded to earlier, graphs are formally defined as a set of vertices V along with the set of edges E between these vertices. In standard fashion, we define a graph as $G = (V, E)$, where $|V| = N$ is the number of nodes in the graph, and $E = N_E$ is the number of edges. We define $A \in R^{NxN}$ as the adjacency matrix related to G (Kipf and Welling, 2016; Defferrard et al., 2017). Fundamentally, graphs are just a way to encode data visually, where properties of graphs represent real elements and concepts within the data. Developing insight into how graphs are used as representations of complex concepts is critical in their efficacy as encoding mechanisms or reasoning over features derived from their structure (Hamilton et al., 2017).

Vertices: In a graph, the objects that are connected are called vertices. These can usually represent entities, which are typically defined as attributes with their relationships and how they are connected to other objects. Given a set of N vertices denoted as V, the i^{th} single vertex we defined as v_i (Hamilton et al., 2017).

Edges: Vertices are connected to one another along edges that characterize the relationships that exist between these vertices. In a strict sense, we defined a single edge between two (not necessarily unique) vertices. Note that a set of N_E edges is denoted as E, and a single edge between the i^{th} and j^{th} vertices is denoted as $e_{i,j}$ (Scarselli et al., 2009; Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017).

Features: In AI, phenomena under study are relegated to quantifiable attributes known as features. Within graph theory for AI, we can utilize these features to express the interactions more deeply between various vertices and edges. In the example of a social network, people are connected to other people, locations, or activities, where features for each person (vertex) could quantify the attributes of a person (e.g., age, popularity, and social media usage) (Gosnell and Broecheler, 2020; Robinson et al., 2015; Needham and Hodler, 2019). Furthermore, features that express relationships between vertices (i.e., edges) could include the quantification of the strength of a relationship or affinity (e.g., familial, colleague, etc.). From a feature standpoint, there can be many considerations per vertex and edge; hence, we represent these as vectors expressed as v_i^F and $e_{i,j}^F$, respectively (Scarselli et al., 2009; Kipf and Welling, 2016).

Neighborhoods: Neighborhoods are smaller portions of a graph made up of nodes and vertices, defined formally as subgraphs, that can be treated as quite distinct sets of vertices and edges. A neighborhood can be iteratively formed through a single vertex by inspecting all connected vertices and edges connected to it. As a neighborhood grows from the ith vertex v_i it will be denoted as the set of neighbor indices $ne[v_i]$. Note that specific criteria can also be defined by specified criteria for the vertex and edge features (Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017; Gosnell and Broecheler, 2020).

States: States are encoded via the information within a given vertices' neighborhood, inclusive of the features and states of the neighborhood's vertex and edge. States are defined as "hidden feature vectors" (Scarselli et al., 2009). In graph theory, these states are iteratively created through a process of extracting features from the previous state's iteration, where classification, regression, or other computation are performed on these iteration states (Kipf and Welling, 2016; Defferrard et al., 2017).

Embeddings: Embeddings are representations acquired through reduction of large feature vectors (Scarselli et al., 2009). The associated vertices and edges within low-dimensional embeddings make it possible to classify them with linearly separable models. The quality of an embedding is measured through the similarity retained in the embedding. Furthermore, these can be "learned" for different parts of the graph (e.g., vertices, edges, neighborhoods, or graphs). Finally, embeddings are also known as representations, encodings, latent vectors, or high-level feature vectors (Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017).

8.4.2 Recurrent graph neural networks (RGNN)

In a standard neural network, successive layers of learned weights work to extract features from an input. After being processed by sequential layers, the resultant high-level features can then be provided to a softmax layer or single neuron for the purpose of classification, regression, etc. A softmax function is often the final neural network activation function that normalizes output of predicted output class probability functions, based on Luce's choice axiom (Luce, 1959). Luce's choice axiom addresses "independence from irrelevant alternatives" (IIA), where the selection of an item over another in a pool of many items is not affected by the existence or non-existence of other items in the pool (Goodfellow et al., 2016). In this same way, the earliest GNN works aimed to extract high-level feature representations from graphs by using successive feature extraction operations (Scarselli et al., 2009), and then fed these high-level features to output functions. The recursive application of a feature extractor, or encoding network, is what provides the RGNN with its name (Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017).

The forward pass: The RGNN forward pass occurs in two main steps. The first step focuses on computing high-level hidden feature vectors for each vertex in the input graph. This computation is performed by a transition function, f. The second step is concerned with processing the hidden feature vectors into useful outputs, using an output function g (Kipf and Welling, 2016).

Transition: The transition process considers the neighborhood of each vertex v_i in a graph, and produces a hidden representation for each of these neighborhoods. Since different vertices in the graph might have different numbers of neighbors, the transition process employs a summation over the neighbors, thus producing a consistently sized vector for each neighborhood. This hidden representation is often referred to as the vertex's state (Scarselli et al., 2009), and it is calculated based on the following quantities (Defferrard et al., 2017; Hamilton et al., 2017; Gosnell and Broecheler, 2020):

(1) v_i^F: the features of the vertex v_i, which the neighborhood is centered around.

(2) $e_{i,j}^F$: the features of the edges which join v_i to its neighbor vertices v_j. Here only direct neighbors are considered, though in practice neighbors further than one edge away may be used. Similarly, for directed graphs, neighbors may or may not be considered based on edge direction (e.g., only outgoing or incoming edges considered as valid neighbor connection).

(3) v_j^F: the features of v_i's neighbors.

(4) h_j^{k-1}: the previous state of v_i's neighbors. Recall that a state simply encodes the information represented in each neighborhood. Formally, the transition function f is used in the recursive calculation of a vertex's k^{th} state as per the following equation:

$$h_i^k = \sum_{j \in ne[v_i]} f\left(v_i^F, e_{i,j}^F, v_j^F, h_j^{k-1}\right),$$

where all h_i^0 are defined upon initialization

We can see that under this formulation, f is well-defined. It accepts four feature vectors, which all have a defined length, regardless of which ver-

tex in the graph is being considered, regardless of the iteration. This means that the transition function can be applied recursively, until a stable state is reached for all vertices in the input graph. If f is a contraction map, Banach's fixed point theorem ensures that the values of h_i^k will converge to stable values exponentially fast, regardless of the initialization of h_i^0 (Khamsi and Kirk, 2001). The iterative passing of messages or states between neighbors to generate an encoding of the graph is what gives this message passing operation its name. In the first iteration, any vertex's state encodes the features of the neighborhood within a single edge. In the second iteration, any vertex's state is an encoding of the features of the neighborhood within two edges away, and so on. This is because the calculation of the k^{th} state relies on the $(k-1)^{th}$ state. To fully elucidate this process, we step through how the transition function is recursively applied. The purpose of repeated applications of the transition function is thus to create discriminative embeddings, which can ultimately be used for downstream machine learning tasks.

Output: The output function is responsible for taking the converged hidden state of a graph $G(V, E)$ and creating a useful and relevant output. Note that the transition function f application to features of $G(V, E)$ ensure all final states $h_i^{k_{max}}$ are encoded in some part of $G(V, E)$. The region size dependency centers around the halting condition (convergence, max time steps, etc.), but often the repeated "message passing" ensures that each vertex's final hidden state has "seen" the entire graph (Scarselli et al., 2009; Kipf and Welling, 2016). These rich encodings typically have lower dimensionality than the graph's input features and can be fed to fully connected layers for the purpose of the ML technique. The output function g, akin to f (the transition function), is implemented by a feedforward neural network (Scarselli et al., 2009), though other means of returning a single value have been used, including mean operations, dummy super nodes, and attention sums (Zhou et al., 2018; Kipf and Welling, 2016; Defferrard et al., 2017; Hamilton et al., 2017). A loss function makes this possible, defined as the error taken from the predicted output and a labeled ground truth (Hamilton et al., 2017). Both f and g can then be trained via backpropagation of errors (Scarselli et al., 2009) for cases that are relevant.

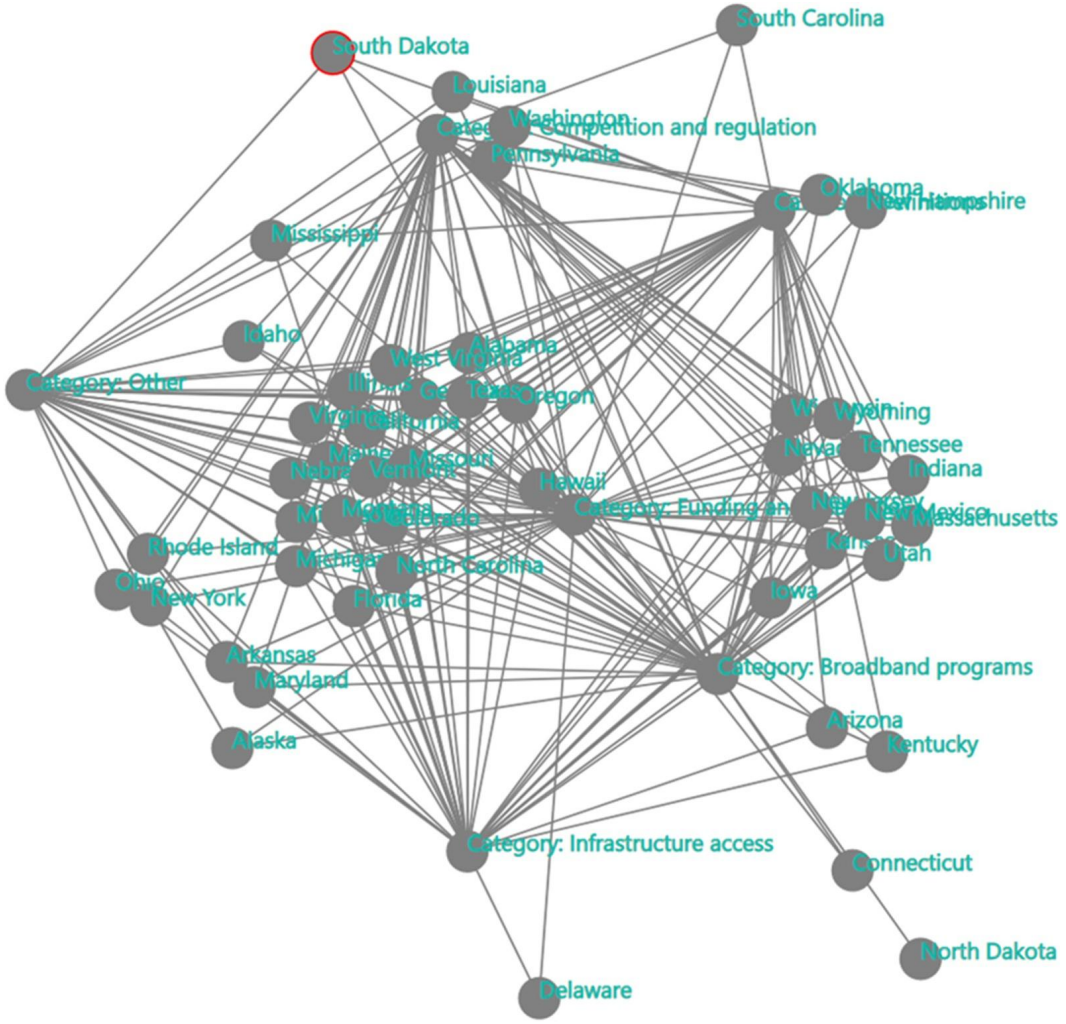


FIGURE 8.5 Graph of Categories and States in the economy of technology dataset.

8.4.3 Economy of technology dataset as a network

The dataset provides unique structural properties when described as a graph. We shall illustrate three different graph views of the dataset we utilized in Section 8.3. We shall illustrate graphs of states with respect to title, category, and topics. Fig. 8.5 illustrates a non-directed graph of Categories and States. Fig. 8.6 presents a graph of states and topics in the data.

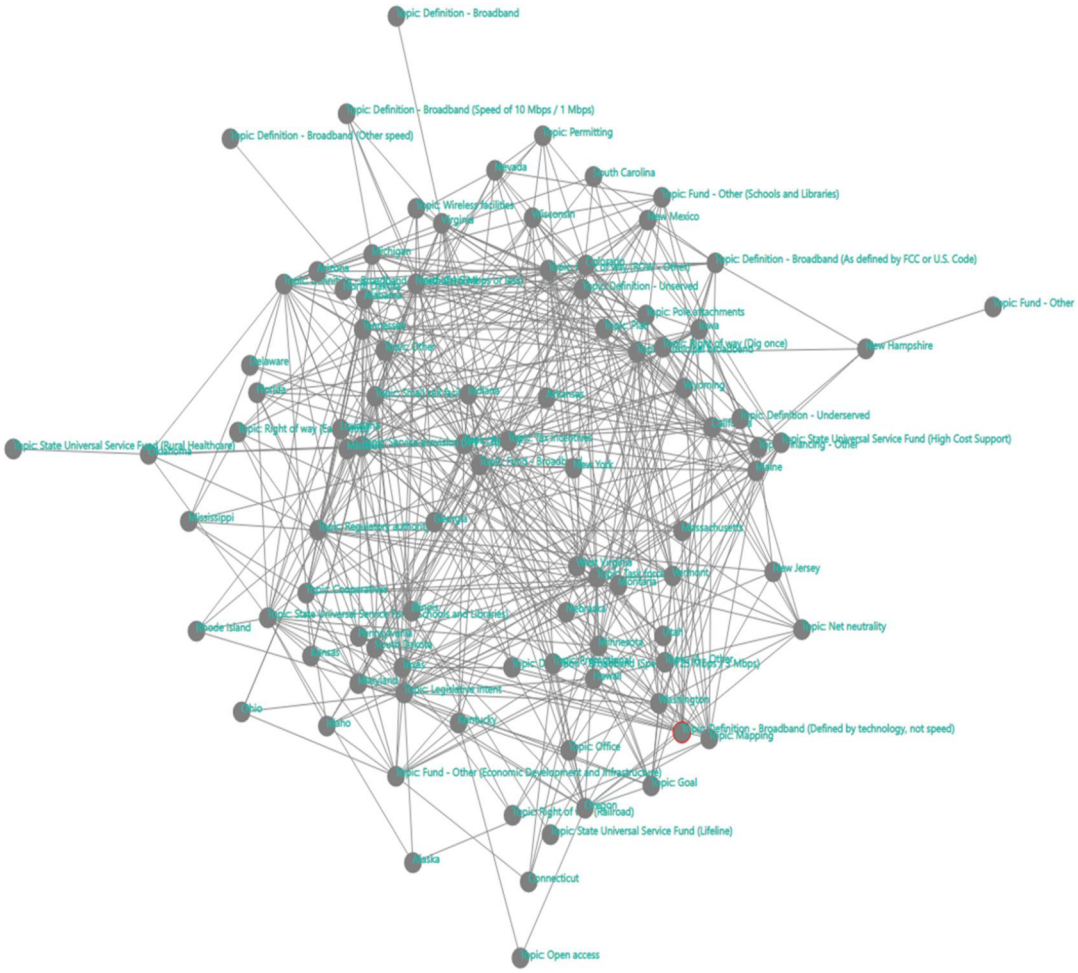


FIGURE 8.6 Graph representation of states and topics in the economy of technology dataset.

Observe that in Fig. 8.5 and Fig. 8.6 the density of connection is far more pronounced than in Fig. 8.7. The graph metric that describes how many edges a vertex has is called centrality, which helps to determine what vertices could be the most important. Also notice some of the vertices only have one edge, which also represents less important vertices. From the standpoint of our dataset, for Fig. 8.5 and Fig. 8.6 that describe topics and categories per state, a graph representation can provide insight into what legal categories and topics are most relevant per state. Note that in for-

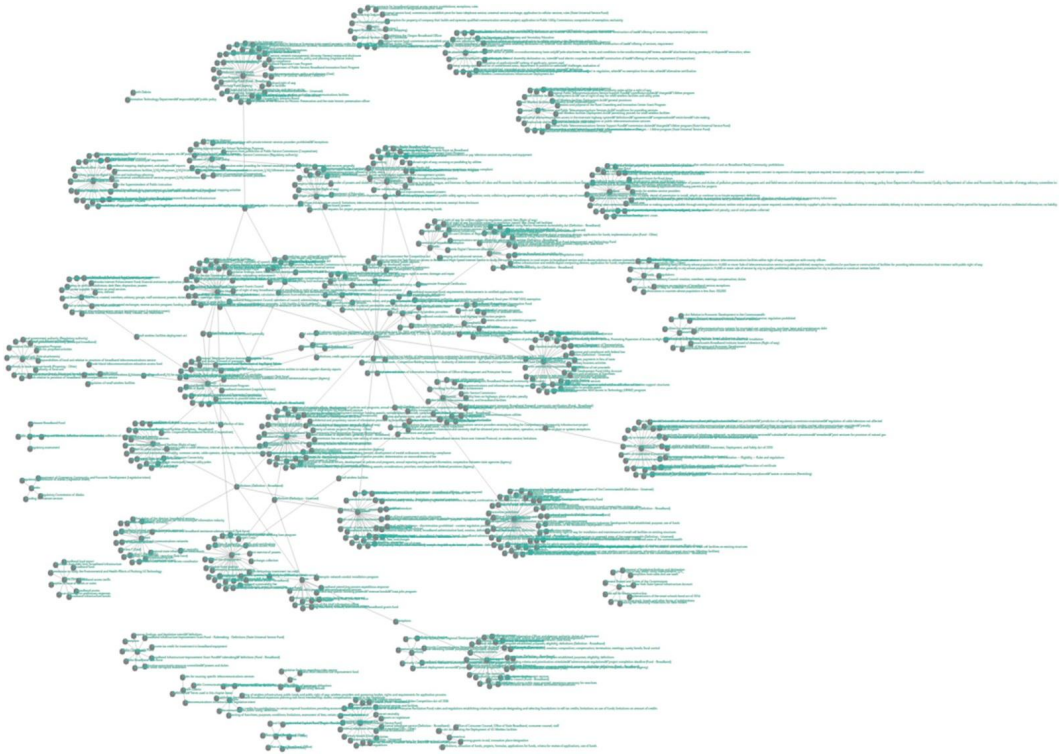


FIGURE 8.7 Graph of U.S. states and titles. Each U.S. state is centered around the related topics that are connected to it. Each cluster represents the U.S. States' respective topics, where lines between clusters identify the relationships that exist.

mulating a GNN to support any type of correlation, anomaly detection, or prediction, such understanding of how to engineer features is critical in formulating an approach toward any type of AI.

8.5 Conclusion

In this chapter, the authors attempted to familiarize the reader with the concepts of causality and its role in AI assurance, and how to build and handle network datasets. We discussed in detail the significance of assuring AI systems applied to the lawmaking process and covered theoretical foundations of causal inference. These concepts were connected through demonstration and explanation of the outcomes of a causal experiment with the economy of technology dataset. We also introduced graph the-

ory, presented examples of structuring the dataset as a network, and elucidated the benefits of such representation. We provided examples of how graph expressions of a sample dataset can provide unique structural insights into the dataset. We hope our work inspired the reader to view their AI-applicable and assurance problems from a new angle and supplied them with helpful background and toolset.

Acknowledgments

The authors would like to thank Dr. Laura Freeman for providing valuable comments on chapter contents and acknowledge Dominick Perini and Amanda Tolman, AI researchers at the A3 research lab (Virginia Tech), for collection and fusion of data used in the experimental part of Section 8.3 of this chapter.

References

H.R.6074 - 116th Congress (2019-2020): Coronavirus Preparedness and Response Supplemental Appropriations Act, 2020. (2020, March 6). https://www.congress.gov/bill/116th-congress/house-bill/6074.

Aldrich, J., 1995. Correlations genuine and spurious in Pearson and Yule. Statistical Science 10 (4), 364–376.

Anuga, A., Nguyen, M., Perini, D., Svetovidov, A., Tolman, A., Wani, Q., Batarseh, F., 2021. Technology policy recommendations using artificial intelligence. In: The International FLAIRS Conference Proceedings, vol. 34.

Batarseh, F., Freeman, L., Huang, C.H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (60), 1–30.

Boire, R., 2018. Understanding AI in a world of big data. Big Data and Information Analytics 3 (1), 22–42.

Bonner, S., Vasile, F., 2018. Causal embeddings for recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems. RecSys '18, pp. 104–112.

Daw, A., Thomas, R., Carey, C., Read, J., Appling, A., Karpatne, A., 2020. Physics-guided architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling. In: SIAM International Conference on Data Mining. SDM '20.

Defferrard, M., Bresson, X., Vandergheynst, P., 2017. Convolutional neural networks on graphs with fast localized spectral filtering. Neural Information Processing Systems 30. arXiv:1606.09375.

Goodfellow, I., Bengio, Y., Courville, A., 2016. 6.2.2.3 Softmax units for multinoulli output distributions. In: Deep Learning. MIT Press. ISBN 978-0-26203561-3, pp. 180–184.

Gosnell, D., Broecheler, M., 2020. The Practitioner's Guide to Graph Data: Applying Graph Thinking and Graph Technologies to Solve Complex Problems. O'Reilly Publishing.

Guo, R., Li, J., Liu, H., 2020. Learning individual causal effects from networked observational data. In: Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20, pp. 232–240.

Hamid, S., 2016. The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare. CUSPE Communications.

Hamilton, W., Ying, R., Leskovec, J., 2017. Inductive representation learning on large graphs (PDF). Neural Information Processing Systems 31. arXiv:1706.02216. via Stanford.

Hartford, J., Lewis, G., Leyton-Brown, K., Taddy, M., 2017. Deep IV: a flexible approach for counterfactual prediction. In: Proceedings of the 34th International Conference on Machine Learning. PMLR 70, pp. 1414–1423.

Hayes-Roth, F., Waterman, D., Lenat, D., 1983. Building Expert Systems. Addison-Wesley Longman Publishing Co., Inc., pp. 89–168.

Hill, J.L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20 (1), 217–240.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), 945–960.

Honderich, T., 1988. A Theory of Determinism: The Mind, Neuroscience, and Life-Hopes. Clarendon Press, Oxford.

Johansson, F.D., Shalit, U., Sontag, D., 2016. Learning representations for counterfactual inference. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. ICML 48, pp. 3020–3029.

Khamsi, M., Kirk, W., 2001. An Introduction to Metric Spaces and Fixed Point Theory. Wiley.

Kipf, T., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations, vol. 5(1), pp. 61–80.

Kobrosly, R., 2020. Causal-curve: a python causal inference package to estimate causal dose-response curves. The Journal of Open Source Software 5 (52), 2523.

Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B., 2019. Meta-learners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences 116 (10), 4156–4165.

Lim, B., Alaa, A., van der Schaar, M., 2018. Forecasting treatment responses over time using recurrent marginal structural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS '18, pp. 7494–7504.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., Welling, M., 2017. Causal effect inference with deep latent-variable models. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, pp. 6449–6459.

Luce, R., 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley, New York. ISBN 978-0-486-44136-8.

Moser, A., Puhan, M.A., Zwahlen, M., 2020. The role of causal inference in health services research I: tasks in health services research. International Journal of Public Health 65 (2), 227–230.

Needham, M., Hodler, A., 2019. Graph Algorithms: Practical Examples in Apache Spark and Neo4j. O'Reilly.
Pearl, J., 2009. Causal inference in statistics: an overview. Statistics Surveys 3, 96–146.
Perry, B., Uuk, R., 2019. AI governance and the policymaking process: key considerations for reducing AI risk. Big Data and Cognitive Computing 3 (2), 26.
Rakesh, V., Guo, R., Moraffah, R., Agarwal, N., Liu, H., 2018. Linked causal variational autoencoder for inferring paired spillover effects. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18, pp. 1679–1682.
Robinson, I., Webber, J., Eifrem, E., 2015. Graph Databases: New Opportunities for Connected Data, 2nd edition. O'Reilly Publishing.
Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2009. The graph neural network model. IEEE Transactions on Neural Networks 20 (1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.
Schwab, P., Linhardt, L., Bauer, S., Buhmann, J.M., Karlen, W., 2020. Learning counterfactual representations for estimating individual dose-response curves. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(04), pp. 5612–5619.
Schwab, P., Linhardt, L., Karlen, W., 2018. Perfect match: a simple method for learning representations for counterfactual inference with neural networks. arXiv preprint. arXiv:1810.00656.
Sekhon, J., 2008. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods. The Oxford Handbook of Political Methodology.
Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning. ICML 70, pp. 3076–3085.
Sharma, A., Kiciman, E., 2020. DoWhy: an end-to-end library for causal inference. arXiv:2011.04216.
Szolovits, P., Sontag, D., 2019. 6.S897 Machine Learning for Healthcare. Massachusetts Institute of Technology: MIT OpenCourseWare. https://ocw.mit.edu. License: Creative Commons BY-NC-SA.
Yao, L., Li, S., Li, Y., Huai, M., Gao, J., Zhang, A., 2018. Representation learning for treatment effect estimation from observational data. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS '18, pp. 2638–2648.
Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Sun, M., 2018. Graph neural networks: a review of methods and applications. arXiv:1812.08434. http://arxiv.org/abs/1812.08434, 2018.
Zuidewijk, A., Chen, Y., Salem, F., 2021. Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. Government Information Quarterly 38 (3).

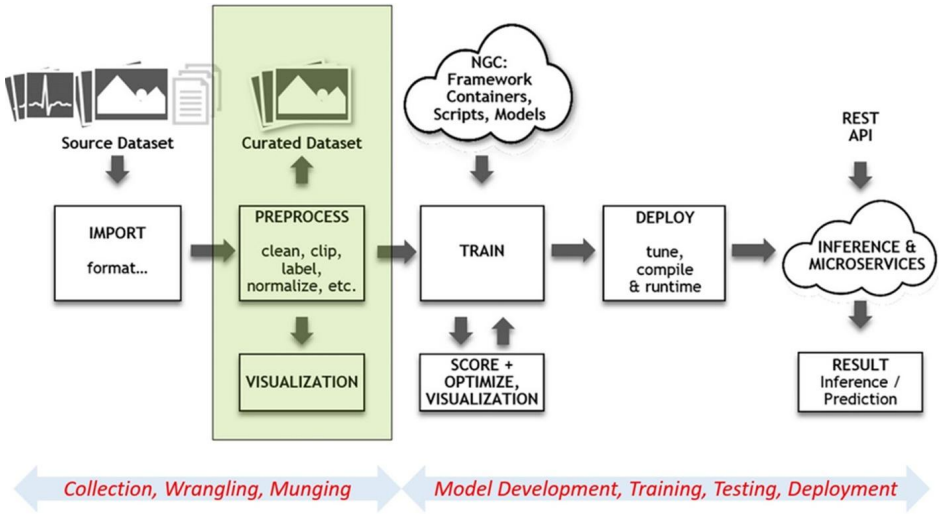
This page intentionally left blank

9

Data collection, wrangling, and pre-processing for AI assurance

Abdul Rahman
Deloitte & Touche, LLP, Baltimore, MD, United States

Graphical abstract



Abstract

Data collection, wrangling, and pre-processing are critical steps within any AI/ML model development lifecycle. These steps precede every model building activity culminating in feature engineering for model formation. This chapter emphasizes the design, development, and implementation of raw data transformation into features in support of AI/ML model development. Integration and preparation of data sets from various sources, such as files, databases, big data storage, sensors or social networks is a key task when you want to build an appropriate analytic model using machine learning or deep learning techniques. Critical to the model building endeavor is the need to have high-quality data

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00022-6

that, unfortunately, has shown to take up 50 to 80 percent of the time for an AI/ML development project. This chapter presents ways to reduce this processing time using data-driven operations "DataOps" (i.e., DevOps for data processing and workflows) pipelines for AI/ML.

Keywords

Data management, data assurance, data science

Highlights

- Understanding the utility of data wrangling, munging, and pre-processing
- Descriptions of data characteristics
- Extract-transform-load (ETL) and extract-load-transform (ELT) processes
- Cleansing strategies that support machine learning and Artificial Intelligence

9.1 Introduction and motivation

Machine Learning (ML) and Artificial Intelligence (AI) algorithms require complete data. In most cases, incomplete data, i.e., missing, or malformed values within a dataset, introduce bias into the predictions that can lead to incorrect or fallacious results. Hence, preparation and treatment of data to reduce or eliminate bias is a critical step in the formation and deployment of reliable and consistent ML and AI. This chapter will present aspects of collection, wrangling, and munging (CWM) of incomplete and impure data to remediate these challenges. CWM focuses on treating raw inoperable data into a state where ML and AI algorithms can consume them. This involves removing ambiguities and impurities within the data through consistent iterative processes.

A first step toward this is performing exploratory data analysis (EDA). EDA precedes most machine learning (ML) and Artificial Intelligence (AI) development, training, testing, validation, and deployment lifecycles to support an evaluation of what is in the data to drive the needed CWM processes. Depending on the use case being developed, most ML pipelines involve a substantial amount of CWM to get the data ready for further steps in the development lifecycle (NVIDIA, 2019).

A large portion of CWM involves addressing missing and malformed data that may arise from missing observations or errors in collection sensors. There are two core techniques for handling missing data: imputation and omission (Hunt, 2017; Luengo et al., 2012). Imputation involves filling in missing data elements, and omission involves removing rows or columns of data that may be malformed or not relevant. Standard imputation techniques (Hunt, 2017) involve four approaches that include replacing data with the following: zeros, the mean, the median, or the mode with respect to observations. Deeper classification of these states, inclusive of techniques to address them, can be asserted through examining conditions by which data is missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR) (van Buuren, 2021). More sophisticated methods involve model-based imputations that predict values based on fitted inference models (Badr, 2019). Treatment of data using any of the above or other techniques should necessarily be documented within the specific data pre- and post-processing (generation) workflows.

Data collection is a major bottleneck in machine learning and an active research topic in multiple communities. Fig. 9.1 depicts the various divisions related to data collection. There are largely two reasons data collection has recently become a critical issue. First, as machine learning is becoming more widely used, we are seeing new applications that do not necessarily have enough labeled data. Second, unlike traditional machine learning, deep learning (DL) techniques automatically generate features, which saves feature engineering costs, but in return may require larger amounts of labeled data. Interestingly, recent research in data collection comes not only from the machine learning, natural language, and computer vision communities, but also from the data management community due to the importance of handling large amounts of data (Roh et al., 2021).

As ML is used in new applications, it is usually the case that there is not enough training data. Traditional applications, such as machine translation or object detection, enjoy massive amounts of training data that have been accumulated for decades. On the other hand, more recent applications have little or no training data. Whenever there is a new product or a new defect to detect, there is little or no training data to start with. The

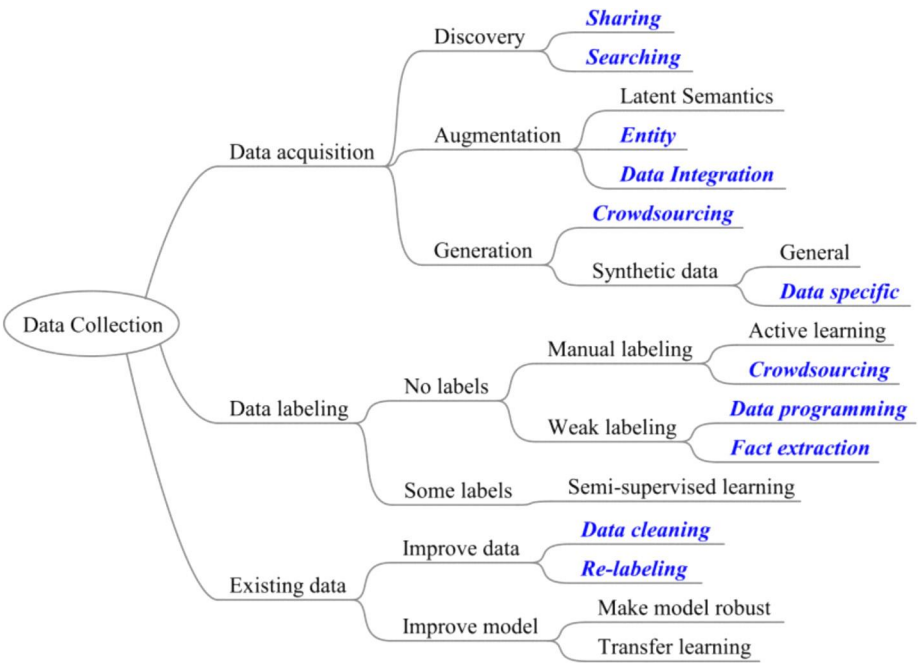


FIGURE 9.1 Data collection depiction for machine learning (Roh et al., 2021).

naive approach of manual labeling may not be feasible, because it is expensive and requires domain expertise. This problem applies to any novel application that benefits from machine learning (Roh et al., 2021).

Moreover, as DL becomes popular, there is even more need for training data. In traditional ML, feature engineering is one of the most challenging steps, where the user needs to understand the application and provide features used for training models. DL, on the other hand, can automatically generate features, which saves us of feature engineering, which is a significant part of data preparation. However, in return, deep learning may require larger amounts of training data to perform well (Bach et al., 2017).

As machine learning needs to be performed on large amounts of training data, data management issues, including how to acquire large datasets, how to perform data labeling at scale, and how to improve the quality of large amounts of existing data become more relevant. Hence, to fully understand the research landscape of data collection, one needs to understand the literature from both the machine learning and data management communities

(Bach et al., 2017). Data collection workflows can be employed to facilitate data acquisition and collection, where these processes are detailed and provide good foundational representations that can be modified in practice (Roh et al., 2021).

As a result, there is a pressing need of accurate and scalable data collection techniques in the era of big data, which is the motivation of this chapter. There are largely three methods for data collection. First, if the goal is to share and search new datasets, then data acquisition techniques can be used to discover, augment, or generate datasets. Second, once the datasets are available, various data labeling techniques can be used to label the individual examples. Finally, instead of labeling new datasets, it may be better to improve existing data or train on top of trained models. These three methods are not necessarily distinct and can be used together. For example, one could search and label more datasets, while improving existing ones. In the next section, we discuss relevant data characteristics that describe data types/formats along with a framework on how to organize data.

9.2 Relevant data characteristics

In general data fall in to four categories that can affect how it is managed: observational, experimental, simulated, or derived/compiled. Prior to developing an AI or ML capability, it is important to work through a series of data management-related steps to help focus the needs for building an optimal capability. It is important to write down a detailed description of how data will be generated or obtained. The situations about when, where, and how much data will be produced drives how the algorithm will consume the data. Also, it is critical to include information on the software that will be used and how the data will be processed.

The lists below capture the possible data formats and types along with some format choices. This list is not exhaustive, but represents some of the data types that can be collected and subsequently used for developing an AI / ML capability.

Formats likely to be accessible in the future are

- Non-proprietary
- Open, with documented standards

- In common usage by the research community
- Using standard character encodings (i.e., ASCII, UTF-8)
- Uncompressed (space permitting) examples of preferred format choices:
- Image: JPEG, JPG-2000, PNG, TIFF
- Text: plain text (TXT), HTML, XML, PDF/A
- Audio: AIFF, WAVE
- Containers: TAR, GZIP, ZIP
- Databases: prefer XML or CSV to native binary formats (DMPTool)

Data that adheres to the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) (Wilkinson et al., 2016) serve to guide data producers as they navigate around data collection challenges. These elements help to maximize the added value gained by contemporary, data sharing, and dissemination methods, processes, and techniques. It is important that the FAIR principles apply not only to "data" in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. The FAIR principles emphasize machine actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention), because humans increasingly rely on computational support to deal with data, given the increase in volume, complexity, and creation speed of data (Wilkinson et al., 2016).

The abbreviation FAIR/O data are sometimes used to indicate that the dataset or database in question complies with the FAIR principles and also carries an explicit data-capable open license. The following principles are worth detailing, as they describe the relevant properties of data taken from (FAIR principles, 2021). The lists below are taken from (FAIR principles, 2021) and (Wilkinson et al., 2016) and present the attributes of each principle of FAIR. These principles are presented here and will be referred to throughout the chapter to both balance and level set the goals and objectives for data collection, wrangling, and pre-processing.

Findable: The first step in (re)using data is to make it "find-able" or "searchable." Metadata and data should be easy to search for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the "FAIRification" process (FAIR principles, 2021).

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource (Azeroual et al., 2018)

Accessible: Once the user finds the required data, they need to know how they can be accessed, possibly, including authentication and authorization.

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
  A1.1 The protocol is open, free, and universally implementable
  A1.2 The protocol allows for an authentication and authorization procedure, where necessary
A2. Metadata are accessible, even when the data are no longer available (Azeroual et al., 2018)

Interoperable: The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

Reusable: The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  R1.1. (Meta)data are released with a clear and accessible data usage license
  R1.2. (Meta)data are associated with detailed provenance
  R1.3. (Meta)data meet domain-relevant community standards

The FAIR principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component). The benefit of referencing data in this manner allows for the ability to incorporate many comprehensive features needed to prepare data to support analytics, modeling, machine learning (ML), and Artificial Intelligence (AI) processes and workflows. These principles will serve as a reference point for core requirements for data collection, wrangling, and pre-processing and will be referenced throughout the remainder of the chapter. The next section will discuss data pre-processing as it pertains to data wrangling and data munging.

9.3 Data pre-processing: data wrangling and munging

The terms "data science," "datafication," "business analytics," and "big data" were coined based on many different developments in data retrieval, storage, and analysis during the last years. Although tools and technologies evolve constantly, understanding and preparing a newly acquired dataset for further usage still requires much time and effort. This initial and very fundamental process of examining and transforming data into a usable form is known as "data wrangling" or "data munging" (Azeroual, 2020; Endel and Piringer, 2015). The data wrangling process involves a broad and deep understanding of the content, structure, and quality issues and necessary transformations as well as appropriate tools and technological resources needed. The whole wrangling procedure needs to be very efficient, especially for small projects or unique datasets, where the effort to automate and document does not seem to be achievable, although necessary. Altogether, data cleaning accounts for 50 percent to 80 percent of the time and costs in analytic or data warehousing projects respectively (Endel and Piringer, 2015).

A key goal of wrangling and munging is to make incremental steps toward improving the overall quality of the data. Data of poor quality, that may have omissions or malformed elements, cannot be properly used in AI/ML algorithms. Fig. 9.2 clearly identifies incorrectness, redundancy, and

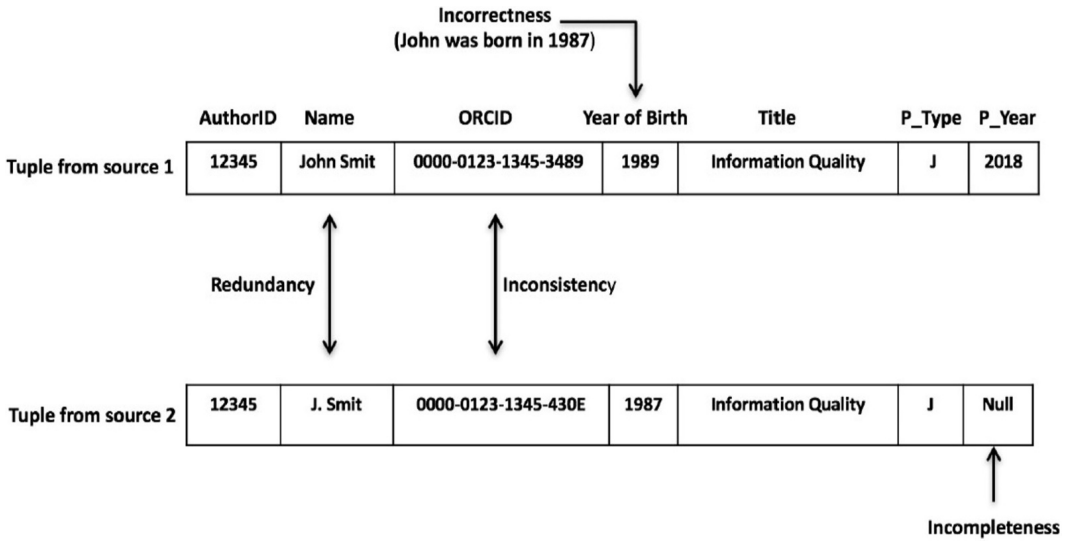


FIGURE 9.2 Examples of data quality problems (Abuosba et al., 2018).

inconsistency issues within the dataset that we will be discussing in the next section, taken from Abuosba et al. (2018). This particular dataset demonstrates a critical need to wrangle and munge the dataset into a format that eliminates or substantially reduces these problems to enable it to be consumed by a to-be developed AI/ML algorithms. Data problems of these types must be addressed way upstream from any subsequent capability and algorithm development.

As discussed in the previous section, these types of data issues can also introduce bias through training, and testing on data of poor quality can prevent an algorithm from operating at its maximum efficiency. This can in turn yield false positives, thereby consuming more "human calories" (i.e., more effort by humans in the loop) than originally intended, and hence nullifying much of the intent of using AI/ML for efficiency and automation in the first place. See Fig. 9.3.

Hence, it is imperative that processing, wrangling, and munging activities align with a core goal of driving towards high enough data quality to be consumed by AI/ML capabilities to support an ongoing and iterative effort to improve algorithm performance and quality. The graphical abstract presents an exemplar workflow to include steps where collection, wran-

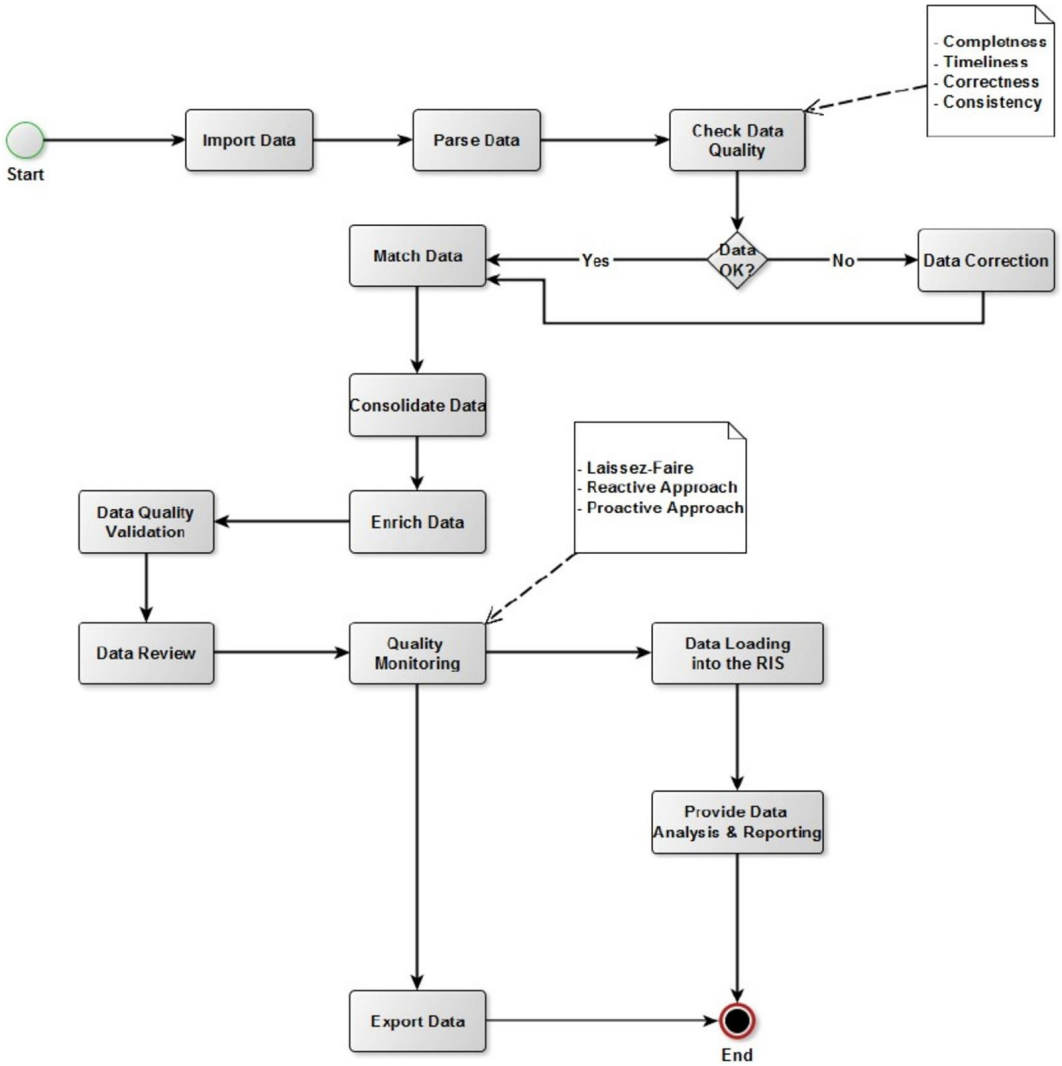


FIGURE 9.3 Exemplar Data Quality Workflow (Abuosba et al., 2018).

gling, and munging are among key steps. In Azeroual et al. (2018) and Azeroual (2020) an example of a data cleansing lifecycle from dirty to cleansed to processed is presented in Figs. 9.4, 9.5, and 9.6.

In Fig. 9.4, notice the **Name** column (or what we will call label) has several records that could be the same, but the first and last name are permuted (i.e., some records have the first name before the last and vice versa). Notice

Author ID	Name	ORCID	Birth Date	Address
12345	John Smit	0000-0123-1345-3487	12/23/1987	123 6th Str Melbourne, 32904
12345	Dr. John Smit	0000-0000-0000-0000	23.12.1987	6th St 32904
12345	John William Smit	0000-0123-1345-3487	872312	10 St 32904 6th

FIGURE 9.4 Example table of unclean data (Abuosba et al., 2018).

Author ID	First	Last	ORCID	Birth Date	Address
12345	John	Smit	0000-0123-1345-3487	1987-12-23	FL; Melbourne; 123 6th St
12345	John	Smit		1987-12-23	FL; Melbourne; 123 6th St
	John	Smit	0000-0123-1345-3487		FL; Melbourne; 123 6th St

FIGURE 9.5 Example of cleansed data (Abuosba et al., 2018).

Author ID	First	Last	ORCID	Birth Date	Address
12345	John	Smit	0000-0123-1345-3487	1987-12-23	FL; Melbourne; 123 6th St
12345	John	Smit	0000-0123-1345-3487	1987-12-23	FL; Melbourne; 123 6th St
12345	John	Smit	0000-0123-1345-3487	1987-12-23	FL; Melbourne; 123 6th St

FIGURE 9.6 Example of cleansed, matched, and consolidated (Abuosba et al., 2018).

also that within the same figure that data with the **Birth Date** label are not uniformly formatted. Both types of data within these labels pose challenges for future AI/ML algorithms that may need to consume these fields. Data wrangling and munging involves formatting these elements in a manner consistent with F, A, and R of the FAIR principles mentioned in the previous section.

With wrangling and munging completed, Fig. 9.5 depicts the cleansed dataset. Notice that the **Name** column has been split into **First** and **Last** to support the FAIR principles of Findability, Interoperability, and Reusability mentioned in the earlier section (Wilkinson et al., 2016). The **Birth Date** and **Address** column are uniformly formatted to support Interoperability and Reusability (Wilkinson et al., 2016). AI/ML algorithms will now benefit from this uniformity as the possible bias manifesting in data loss introduced through poor formatting has been mitigated through these data wrangling and munging steps.

In Fig. 9.6, the data within the cleansed table in Fig. 9.5 is matched and consolidated to support the FAIR principles of Interoperability and Reusability (Wilkinson et al., 2016).

In the next section, ETL and ELT data architectures will be presented.

9.4 Data processing architectures: ETL & ELT

In Marín-Ortega et al. (2014), the authors compare and contract extract-transform-load versus extract-load-transform architectures.

In a typical data architecture, to include business intelligence (BI) data warehouses (DW) infrastructures, data is extracted from ingested sources, firstly transformed, then cleaned and loaded (Jorg and Dessloch, 2009). Before data are loaded into a DW for example, it is necessary to process "raw data" for a variety of reasons. Incoming data must be normalized. Also, the source data may contain erroneous, corrupted or missed data, so the process of cleaning and re-consolidation are needed. This also presents the case for imputation as discussed earlier (Hunt, 2017; Luengo et al., 2012). This pre-processing is commonly known as extract, transform and load (ETL): data are first extracted from the original data source, then transformed, including normalization and cleansing, and finally loaded into the DW (Jorg and Dessloch, 2008). While database technologies used for data warehousing had seen tremendous performance and scalability enhancements over the past decade, ETL has not been improved in scalability and performance in the same level of degree as database. As a result, most infrastructures are increasingly experiencing a bottleneck: data cannot be easily acquired with necessary actuality. Clearly, to provide near real-time capabilities, this bottleneck needs to be resolved. Costs of data storage were always a significant factor, but with the advent of the cloud and reduced hardware costs, this is becoming cheaper with time. As a result, analysis can be performed over larger datasets with smaller investments. Furthermore, former (but robust) extract, transform and load approaches cannot be easily applied to answer all needs of business, which includes working with big data and, as a result, new approaches and/or architectural changes are needed. The main disadvantage of ETL is that data must be firstly transformed and only then loaded. It means that during the trans-

formation phase, mass amounts of potentially valuable data are thrown away (Jorg and Dessloch, 2009). However, to eliminate drawbacks of ETL, improvement of latest storage techniques can be used (Jorg and Dessloch, 2008).

ELT (extract, load and transform) in comparison with ETL, has four following advantages: (1) flexibility in adding new data sources (EL part); (2) aggregation can be applied multiple times on same raw data (T part); (3) transformation process can be re-adopted, even on legacy data; (4) speed-up process of implementation. Also, transformation with ELT can be applied and re-applied taking into account changes in business requirements. Based on above reasons it is more preferable to adopt extract, load and transform (ELT) instead of extract, transform and load (ETL) in the within data architectures. The next section discusses data operations automation management known as DataOps.

9.5 DataOps: data operations automation management

DevSecOps pipeline design and development to support data wrangling and pre-processing, automation best practices for pre-processing steps to include error handling DevOps (Capizzi et al., 2020; Lwakatare et al., 2015) is an approach for software development and (IT) system operation combining best practices from both such domains to improve the overall quality of the software-system, while reducing costs and shortening time-to-market. The DevOps formalism can be generalized as a good practice for improving a generic product or service development and operation, by connecting these through feedback from operation to development. An important feature of DevOps is the automation of such a process: continuous delivery (CD) enables organizations to deliver new features quickly and incrementally by implementing a row of changes into the production via an automated assembly line, called the continuous delivery pipeline. This is coupled with continuous integration (CI) that aims at automating the software/product integration process of code, modules and components, thus identifying a CI/CD pipeline. The tools adopted to implement this high degree of automation in the DevOps process identifies a toolchain. DevOps toolchain tools are usually encapsulated into different, independent con-

tainers deployed into physical or virtual servers (typically on Cloud), and then managed by specific scripts and/or tools (e.g. Jenkins), to orchestrate and coordinate them automatically. Such DevOps principles have been therefore either specialized to some specific software/application domains (security: SecOps, SecDevOps, DevSecOps (Lwakatare et al., 2015), system administration: SysOps, Web - WebOps or WebDevOps) or even adopted, rethought, and adapted in other contexts, such as Artificial Intelligence (AIOps) and machine learning (MLOps), and data management (DataOps).

The latter, DataOps, aims at mainly organizing data management according to DevOps principles and best practices. To this end, DataOps introduces the concept of data flow pipeline and toolchain, which is to be deployed in containerized (Cloud) environment providing feedback on performance and QoS of the overall data management process, used to real-time tune the pipeline to actual operational needs and requirements. As discussed above, DevOps pipeline automation involves different toolchains, each continuously generating messages, logs, and data, including artifacts. To achieve DevOps aims and goals, such data has to be properly managed, collected, processed, and stored to provide insights from operations to the development stages. DevOps data management could therefore be quite challenging, due to the large amount of data to be considered, and its volume and variety (Capizzi et al., 2020; Lwakatare et al., 2015). In the next section data tagging, provenance, and lineage will be discussed.

9.6 Data tagging, provenance, and lineage

Upon completion of CWM processes, data may be tagged to identify its source, how it was acquired, and who it can be shared with. These types of metadata are critical in tracing each element of a data source that may be used for information sharing. Within the national intelligence enterprise, this type of sharing supports not only the aggregation of information assurance metadata (including enterprise data headers), but also allows inter-agency access control, automated exchanges, and appropriate protection of shared intelligence. Using agreed upon data header standards, a structured, verifiable representation of security metadata bound to the intelligence data enables data systems to become inherently “smarter” about

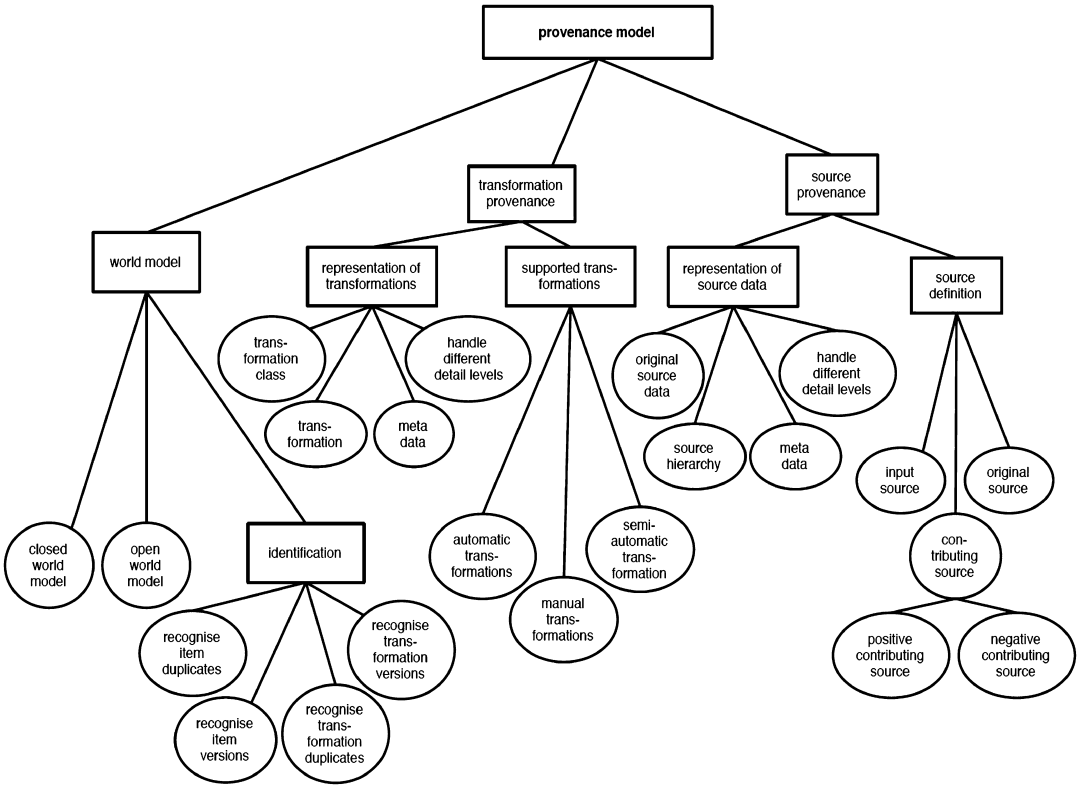


FIGURE 9.7 Data provenance conceptual model (Glavic and Dittrich, 2007).

the information flowing in and around it. Such a representation, when implemented with other data formats, improves user interfaces and data processing utilities, and can provide a larger more robust information assurance infrastructure capable of automating some of the more critical but menial management and exchange processes (DNI, 2021).

Information sharing largely depends on data provenance and lineage, that is, documenting where a piece of data came from and the process by which it arrived at its storage target. Fig. 9.7 presents this. As described in the figure, this is becoming increasingly important, especially in scientific databases, where understanding provenance is crucial to the accuracy and currency of data (Buneman et al., 2001; Simmhan et al., 2005; Glavic and Dittrich, 2007). Extending these use cases to other venues, such as national security, involves application of similar principles.

Data lineage enables an end-to-end data-centric audit trail that can facilitate all levels and forms of compliance. As assuring AI can center on meeting certain compliance standards, data lineage focuses on tracing data quality issues along with other errors to root causes, where impact analysis can be conducted on any proposed change. Physically and logically separated systems can benefit from data lineage by describing various metadata connectivity constructs, where identification of business rule discrepancies and data incompleteness is critical. This also supports other governance participants to respond to issues before they become a problem, making possible data quality improvement for the possible reuse of existing information in all its forms.

The intelligence community (IC) has standardized the various classification and control markings established for information sharing within the information security markings (ISM), need-to-know (NTK), and access, rights, and handling (ARH) XML specifications of the Intelligence Community Enterprise Architecture (ICEA) data standards. The IC enterprise data header XML specification further expands on this body of work, adapting and extending it as necessary to meet mission-unique needs. By specifying a data object's header information required for exchange on the IC enterprise, EDH ensures a secure method of information sharing and discovery, supporting use cases, such as the IC Cloud.

Though increases in data volumes can lead to improved performance and increased accuracy of models, what is more important is the quality of the data. Training models on data that is too noisy or lacks variance can lead to models that lack real predictive power. As a result, the assurance of AI models involves ensuring that data used to train them is of sufficient quality to promote maximal predictive power, while avoiding setbacks, such as overfitting.

References

Abuosba, Mohammad, Azeroual, Otmane, Saake, Gunter, 2018. Data quality measures and data cleansing for research information systems. In: 8th Vienna International Conference on Mathematical Modelling. Journal of Digital Information Management, Digital Information 16 (1), 12–21.

Azeroual, Otmane, 2020. Data wrangling in database systems: purging of dirty data. Data 5, 2.

Azeroual, Otmane, Saake, G., Schallehn, E., 2018. Analyzing data quality issues in research information systems via data profiling. International Journal of Information Management 41, 50–56.

Bach, Stephen H., et al., 2017. Learning the structure of generative models without labeled data. In: Precup, Doina, Teh, Yee Whye (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70. PMLR, pp. 273–282.

Badr, Will, 2019. 6 different ways to compensate for missing values in a dataset (data imputation with examples). Retrieved from: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

Buneman, Peter, Khanna, Sanjeev, Wang-Chiew, Tan, 2001. Why and where: a characterization of data provenance. In: Van den Bussche, Jan, Vianu, Victor (Eds.), Database Theory - ICDT 2001. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 316–330.

Capizzi, Antonio, Distefano, Salvatore, Mazzara, Manuel, 2020. From DevOps to DevDataOps: data management in DevOps processes. In: Bruel, Jean-Michel, Mazzara, Manuel, Meyer, Bertrand (Eds.), Software Engineering Aspects of Continuous Development and New Paradigms of Software Production and Deployment. Springer International Publishing, Cham, pp. 52–62.

Endel, Florian, Piringer, Harald, 2015. Data wrangling: making data useful again. In: 8th Vienna International Conference on Mathematical Modelling. IFAC-PapersOnLine 48 (1), 111–112.

go-fair.org, FAIR principles, 2021. Retrieved from: https://www.go-fair.org.

Glavic, B., Dittrich, K., 2007. Data provenance: a Cctegorization of existing approaches. In: Kemper, A., et al. (Eds.), GI-Edition - Lecture Notes in Informatics (LNI). Proceedings 103. Gesellschaft für Informatik (GI), Bonn, pp. 227–241.

Hunt, L.A., 2017. Missing data imputation and its effect on the accuracy of classification. In: Palumbo, F., Montanari, A., Vichi, M. (Eds.), Data Science. Springer International Publishing, Cham, pp. 3–14.

Jorg, T., Dessloch, S., 2008. Towards generating ETL processes for incremental loading. In: Proc. of Int. Symposium on Database Engineering and Applications (IDEAS), pp. 101–110.

Jorg, T., Dessloch, S., 2009. Formalizing ETL jobs for incremental loading of data warehouses. Fachtagung des GI-Fachbereichs Datenbanken und Informationssysteme 13, 327–346.

Luengo, J., García, S., Herrera, F., 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems 32, 77–108.

Lwakatare, Lucy Ellen, Kuvaja, Pasi, Oivo, Markku, 2015. Dimensions of DevOps. In: Lassenius, Casper, Dingsøyr, Torgeir, Paasivaara, Maria (Eds.), Agile Processes in Software Engineering and Extreme Programming. Springer International Publishing, Cham, pp. 212–217.

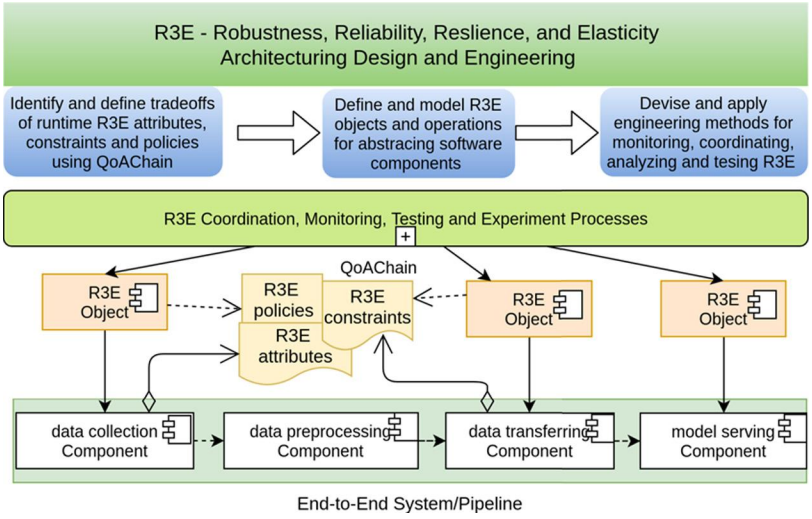
Marín-Ortega, Pablo Michel, et al., 2014. ELTA: new approach in designing business intelligence solutions in era of big data. Procedia Technology 16, 667–674.
NVIDIA AI, 2019. Accelerating the entire deep learning pipeline. Retrieved from: https://medium.com/@NvidiaAI/accelerating-the-entire-deep-learning-pipeline-bff0e4e05fc.
Office of the Director of National Intelligence (DNI), 2021. IC-enterprise data header. Retrieved from: https://www.dni.gov/.
Roh, Yuji, Heo, Geon, Whang, Steven Euijong, 2021. A survey on data collection for machine learning: a big data - AI integration perspective. IEEE Transactions on Knowledge and Data Engineering 33 (4), 1328–1347.
Simmhan, Yogesh L., Plale, Beth, Gannon, Dennis, 2005. A survey of data provenance in E-science. SIGMOD Record 34 (3), 31–36.
van Buuren, Stef, 2021. Flexible Imputation of Missing Data, 2nd edition. Chapman & Hall/CRC Interdisciplinary Statistics.
Wilkinson, Mark D., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3 (160018).

10

Coordination-aware assurance for end-to-end machine learning systems: the R3E approach

Hong-Linh Truong
Department of Computer Science, Aalto University, Espoo, Finland

Graphical abstract



Abstract

Concerns of robustness, reliability, resilience, and elasticity in Machine Learning (ML) systems are important, and they must be considered in trade-off with efficiency factors. However, they need to be supported and optimized in an end-to-end manner, not just for ML models. In this chapter we present a conceptual approach to architectural design and engineering of the robustness, reliability, resilience, and elasticity (R3E) for end-to-end big data ML systems at runtime.

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00024-X

*We propose quality of analytics as a contractual means for optimizing end-to-end big data machine learning (*BDML*) systems. Based on that, we propose to define and abstract diverse types of components under R3E objects and devise operations and metrics for managing R3E attributes. Through a set of proposed coordination, monitoring, analytics, and testing methods, we identify essential tasks for tackling R3E concerns when developing* BDML *systems. Finally, we illustrate our approach with an example of an end-to-end* BDML *system for building objects classifications.*

Keywords

Software systems, machine learning, big data, software architecture, elasticity, cloud computing, engineering analytics

Highlights

- This chapter characterizes robustness, reliability, resilience, and elasticity (R3E) in architectural designs for end-to-end big data machine learning systems.
- We provide a novel model of quality of analytics chain (QoAChain) to abstract and define constraints for assuring robustness, reliability, resilience, and elasticity of end-to-end machine learning.
- We present a concept of R3E objects and operations abstracting components in big data machine learning systems.
- We discuss engineering methods for coordinating, monitoring, analyzing, and testing R3E attributes.

10.1 Introduction

Big data machine learning (*BDML*) systems enable different types of ML-based pipelines, which deal with big data in motion or at rest. End-to-end *BDML* systems support tasks from processing raw data to producing inference results. Thus *BDML* systems involve several different software components, including data sources collectors/connectors, message brokers, edge data preprocessing and aggregators, cloud data stores, ML serving platforms, and ML services. These components are cross-layered and cross-infrastructural, due to the nature of diverse ML pipelines and data to be supported by such systems. Therefore, components of an end-to-end

BDML system are potentially deployed and offered in multiple edge and cloud infrastructures. Typically, the data to be inferenced and the application using the ML model-as-a-service are from the consumer, whereas the ML model-as-a-service can be run in the edge or cloud by the ML service provider, which offers the service to many consumers. Furthermore, computing, storage, and communication services might be offered by other providers. Such systems for real-world ML must be designed with robustness, reliability, resilience, and elasticity (R3E) concerns from a multi-party perspective. Although individual components may be designed and tested with certain degrees of R3E, the challenging question for the development of end-to-end *BDML* systems is to guarantee expected runtime R3E attributes across layers and infrastructures. Therefore, recently, the role of software systems and underlying distributed computing platforms and their intersections with ML have been discussed intensively. Since *BDML* systems are complex and typically used for critical businesses, the R3E attributes play a key role in *BDML* software architectures and implementations. Ensuring R3E is challenging for complex software systems, because R3E attributes are highly interdependent and multi-dimensional. Especially, R3E attributes in *BDML* systems are related to three aspects: *services, data,* and *ML models*. The key research question in our work is how to build and optimize the R3E attributes for *BDML* systems in an end-to-end manner; and the "end-to-end" aspect forces us to examine various components of *BDML* systems together, following their dependencies, interactions and functions.

This chapter presents a novel conceptual approach to R3E engineering for *BDML* systems, in which we will focus on software architecture and design aspects. We develop abstractions and methods for architectural designs, runtime optimization, and engineering analytics in *BDML*. Our approach considers different levels of abstractions of *BDML*, from data collection to training to model serving to determine key constraints, engineering steps, monitoring, and management processes for making *BDML* robust, reliable, resilient, and elastic. Our conceptual approach makes the following contributions:

- QoAChain as a means to combine quality of analytics constraints, services contracts, and data contracts for specifying runtime R3E attributes and constraints
- abstractions and models for R3E objects and operations in *BDML*
- engineering methods for coordinating, monitoring, analyzing, and testing R3E attributes

Our approach covers key aspects of R3E engineering and layouts the foundational work for the development of specific techniques and tools to support R3E in *BDML* systems. To illustrate our approach, we will use a realistic example of end-to-end *BDML* for building objects classifications.

The rest of this chapter is organized as follows: Section 10.2 characterizes R3E in *BDML* and presents our motivating examples and research questions. Section 10.3 presents our R3E approach. We present a concrete example of R3E aspects and identified requirements in Section 10.4. Further related work will be discussed in Section 10.5. We conclude the chapter and outline the future work in Section 10.6.

10.2 Background and motivation

10.2.1 Background – characterizing *BDML*

A *BDML* system can be characterized as follows:

- *system structures and functions*: a *BDML* system includes various components implementing different functions. Examples of components are a data storage service and an ML serving platform, whose functions are storing data and serving ML models, respectively. Components have different relationships and possible inputs and outputs. R3E attributes can be associated with individual components, a set of components, and the system as a whole.
- *supporting computing, data, and communication infrastructures*: computing infrastructures provide different types of computing resources for different tasks, notably data preprocessing, training, and serving. Typically, such infrastructures include advanced computing systems, such as CPU/GPU resources, containers and Kubernetes, message brokers, and edge systems. The data infrastructures provide data for training and data

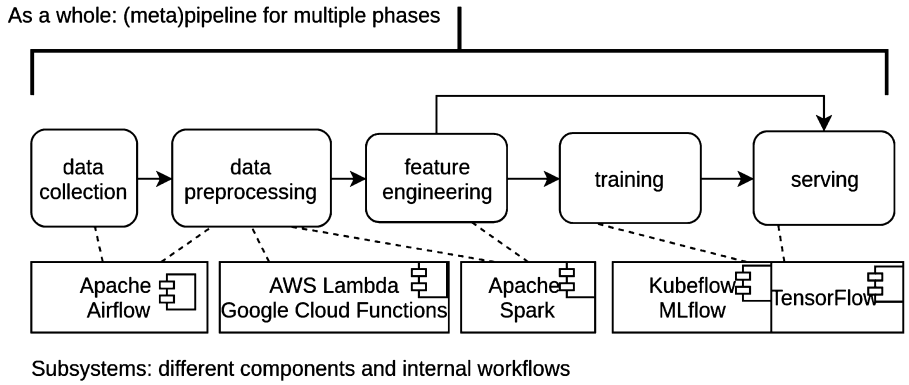


FIGURE 10.1 Example of ML pipelines and components in *BDML* systems.

being inferenced as well as other types of data related to ML, such as ML model experiments and performance of ML services. Our focus is on big data infrastructures.

- *runtime quality/capabilities*: they include multiple attributes, for example, regarding to fault-tolerance, high performance, high availability, and security of software services. From the data view, a *BDML* system has to deal with big data characteristics, such as volume, variety, velocity, and veracity, from the data source to the end of the ML pipelines. Furthermore, ML models have different quality attributes, depending on domain requirements and business contexts.

For example, Fig. 10.1 shows a view from common tasks in end-to-end *BDML* pipelines. From the ML pipelines perspective, the system-as-a-whole can be seen as a meta pipeline orchestrating different sub systems, where each subsystem can be implemented differently, such as with Airflow, Lambda, TensorFlow, and other supporting services. Each of them requires a variety of components for data, software services, ML algorithms, and pipeline orchestration.

In this chapter, we consider well-studied R3E attributes in the state-of-the-art literature:

- robustness attribute (Gribble, 2001; Laranjeiro et al., 2021) is about the ability to cope with errors, such as with the error of the data (Sehwag et al., 2019).

- reliability attribute (Littlewood and Strigini, 2000; Saria and Subbaswamy, 2019; Elsayed, 2012) is about the ability to properly function/operate according to the service specification, e.g., the availability of a service must be 99%.
- resilience attribute (Trivedi et al., 2009; Brtis et al., 2021) is about the ability to hold out required capabilities under adversity, e.g., due to system failures or security attacks.
- elasticity attribute (Dustdar et al., 2011) is about the ability to stretch and return to normal service capabilities, e.g., under external forces of usage demands.

We will rely on common definitions and usages of these attributes from the big data and ML perspectives. Table 10.1 gives examples about key R3E concerns and factors from big data and ML views of *BDML* systems. Our approach will support R3E attributes in such common senses.

10.2.2 Motivating example: machine learning for classifying building elements

We consider a prototype for an end-to-end *BDML* system for classification of building information modeling (BIM) elements in the architect, engineering, and construction domain. ML-based BIM classification allows to speedup the design and check conformity of building models. In our collaboration, an initial end-to-end BIM *BDML* system has been developed using various AWS services for moving data; ML capabilities are built with TensorFlow and Keras (Ryu et al., 2021). Fig. 10.2 shows a simplified view of a new architectural design for the discussion of the role of R3E in this chapter, where we are leveraging serverless platforms to better manage and optimize the complex relationships between various components. In the new design, data exported from user tools will be moved to *Data Service*. New data will be detected, and *preprocessing* and *feature engineering* will be triggered by serverless platforms, before *ML Service* serves requests of classifications. Atop the *ML Serving Platform*, we have *ML Service* with different *ML Models* for BIM classification.

Fig. 10.2 not only shows an end-to-end system with various components, but also creates clear interfaces between different sub-pipelines, such as

Table 10.1 Common R3E with big data and ML concerns.

R3E attributes	Cases from big data view	Cases from machine learning view
Robustness	deal with erroneous and bad data (Zhang et al., 2017), data processing job robustness	dealing with imbalanced data, learning in an open-world (out of distribution) situations (Kulkarni et al., 2020; Sehwag et al., 2019; Saria and Subbaswamy, 2019; Hendrycks and Dietterich, 2019)
Reliability	reliable data sources, support of quality of data (Zhang et al., 2020; Lee, 2019), reliable data services (Kleppmann, 2016), reliable data processing workflows/tasks (Zheng et al., 2017)	reliable learning and reliable inference in terms of accuracy and reproducibility of ML models (Saria and Subbaswamy, 2019; Henderson et al., 2017); uncertainties/confidence in inferences; reliable ML service serving
Resilience	software bugs, infrastructural resource failures, fault-tolerance and replication for data services and processing (Yang et al., 2017)	bias in data, adversary attacks in ML (Katzir and Elovici, 2018), resilience learning (Fischer et al., 2018), computational Byzantine failures (Blanchard et al., 2017)
Elasticity	utilizing different data resources; increasing and decreasing data usage with respect to data volume, velocity, and quality; elasticity of underling resources for data processing (Wang and Balazinska, 2017)	elasticity of resources for computing (Huang et al., 2015; Harlap et al., 2017; Gujarati et al., 2017), elasticity of model parameters; performance loss versus model accuracy; elastic model services for performance

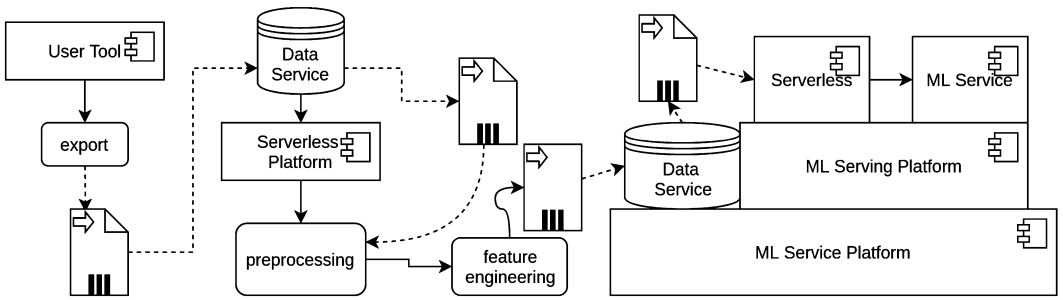


FIGURE 10.2 Overview of an end-to-end big data machine learning system for BIM.

preprocessing, *feature engineering*, and *serving*, enabling us to carry out different performance/cost optimizations for different pipelines and their un-

derlying components. It also allows us to deal with R3E attributes that are more flexible for subpipelines and underlying components. However, it creates various R3E concerns that need to be addressed together. For example, *ML Service* has to be elastic to support different requirements with respect to accuracy, cost, and performance. This is dependent on the elasticity of the underlying *ML Service Platform*, which is strongly linked to computing resources, and on the output of *feature engineering*, which in turn is strongly dependent on the exported data sent to *Data Service* and *preprocessing* robustness and reliability.

10.2.3 Research questions

Key issues of end-to-end *BDML* are not just about efficiency, such as highly responsive in serving the classification with a minimum cost, but also in trade-offs with robustness, reliability, resilience, and elasticity. For example, in the BIM scenario (Section 10.2.2), the accuracy of inference results from *ML service* and the resilience of *ML service* are more important than the response time due to the business nature of the domain. As recognized in (Ackley, 2013), performance must also be aligned with robustness and resilience. The robustness of the ML model depends on the data input. From the computation and network, reliability concerns for edge-cloud have many issues (Suryavansh et al., 2019; Nguyen et al., 2019). The reliability concern from the data aspect is that the data source must provide "reliable data," interpreted as the data quality and quantity satisfied the required conditions. On the other hand, from the service viewpoint, the ML model serving will be considered as a reliable service when it can return the results in specified time. This turns out to be dependent on multiple factors, such as the reliability of the underlying computing resources (e.g., no failure) and the elasticity of the resources (e.g., to assure response times in the expected range).

We see that R3E concerns exist in different parts of a *BDML* system. However, currently, there is no systematically way to capture, represent, monitor, and optimize such R3E attributes from the design and architecture viewpoint. Our vision in this chapter is the following:

R3E attributes can be systematically modeled, programmed, and captured at different levels of abstractions in *BDML* systems, enabling the coordinated optimization of these attributes in an end-to-end view, based on specific contexts of the intended end-to-end ML pipelines executed in *BDML* systems.

Consequently, we have the following important research questions (RQs):

- RQ1: *what would be the model for abstracting R3E constraints?* With diverse R3E concerns, we need to capture key R3E attributes and describe them into appropriate constraints.
- RQ2: *how can we abstract complex components in the R3E view and define suitable operations for managing R3E?* Components in *BDML* systems need to be managed through the R3E view, which should capture attributes and essential operations.
- RQ3: *which are the key engineering methods for achieving R3E?* Engineering methods for monitoring and managing R3E attributes across components of *BDML* systems must be laid out, paving the way to develop suitable tools and frameworks.

The R3E approach will provide key conceptual steps and components to address the above-mentioned questions.

10.3 Key elements of R3E approach

10.3.1 QoAChain: chaining diverse types of quality constraints as a contract for optimizing end-to-end *BDML*

RQ1 requires us to determine how to abstract R3E attributes and to specify R3E concerns for optimizing *BDML* systems. Fig. 10.3 gives a high-level view of the complex relationships among various concerns of different attributes when optimizing *BDML*. Given an application, big data and ML pipelines are combined and executed to analyze input data (data in) and produce results. Such executions are carried out with edge-cloud resources as services. There are many questions with respect to attributes associ-

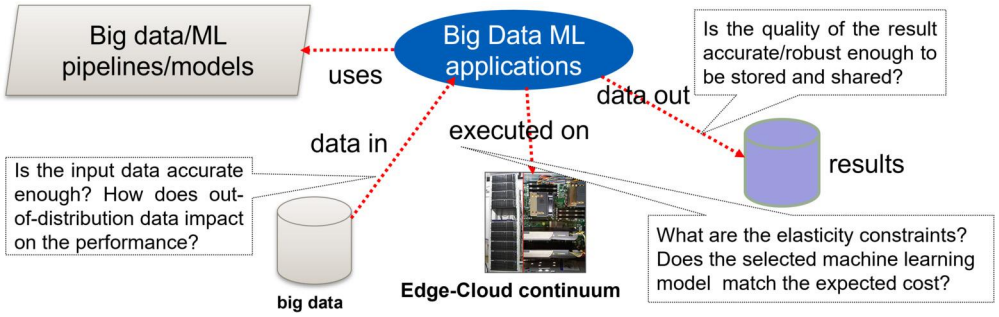


FIGURE 10.3 Concerns among components and stakeholders in optimizing ML.

ated with used models, input data, results, and execution environments of computing, data and communication services, as exemplified in Fig. 10.3. These concerns are from different involved stakeholders, such as the application users, the *BDML* system provider, the developer and scientist of ML pipelines, and the resources provider. Overall, they reflect the concerns of dealing with trade-offs between R3E and efficiency.

Consider the complex relationships among various components and stakeholders in *BDML*; we choose to combine the concept of quality of analytics (QoA) (Truong et al., 2018), machine learning service contracts (Truong and Nguyen, 2021), and data contracts (Balint and Truong, 2017; Truong et al., 2012) for end-to-end *BDML*. We summarize these works in the following:

- Quality of analytics (QoA) (Truong et al., 2018) emphasizes the need to optimize data analytics based on specific contexts that is elastic. It characterizes complex relationships between quality of results, performance, and cost that are not fixed, but changing according to requirements, even for the same system:
  - Quality of results, outputted from data analysis tasks, including ML ones, are characterized by the user/domain expert, e.g., quality of data of the output and the accuracy of predictions.
  - Input data has complex characteristics with respect to, for example, quality of data and data volume and velocity, that strongly influence infrastructural resources as services, such as task execution, computing machines, and storage.

  - Complex types of cost (money) and performance are based on business purposes, contextually expected and changed by involved stakeholders.
- The recent work on machine learning contracts (Truong and Nguyen, 2021) defines contractual terms between ML service providers and ML customers. ML contracts focus on ML-specific attributes, such as inference accuracy, at runtime that are agreed between the customers and the services.
- Existing data contracts (Truong et al., 2012; Balint and Truong, 2017) focus on constraints on data to be delivered from data sources (providers) to consumers. They focus very much on quality of data attributes.

Clearly the above-mentioned concepts aim at guaranteeing important constraints seen in *BDML* systems. ML-specific attributes, data quality attributes, and common service attributes can be associated with various parts of a *BDML* system. The associations can be for individual components or a whole pipeline, and can indicate different expectations in the *BDML* system. Due to the diversity of component types, inputs and outputs, it is difficult to have a single way to specify such constraints for *BDML* systems.

A "reliable *BDML* system" should guarantee the specified runtime quality attributes built from the work on QoA, ML contracts, and data contracts, while maintaining designed R3E attributes. Due to the nature of ML systems, we can use these concepts to specify constraints for different parts of a *BDML* system for different purposes, such as:

- data contract: a constraint on data completeness for input IoT data
- ML contract: a constraint on inference accuracy for inference results
- common service contract: a constraint on the response time for the end-to-end processing

These examples show that runtime constraints can be defined for different components for different attributes, and these constraints might be specified by different models. A *BDML* system is designed and optimized for different ML pipelines, which serve different business purposes, depending on the usage of the resulting outcome of the pipelines and the business goal of the provider of the pipelines supported by the *BDML* sys-

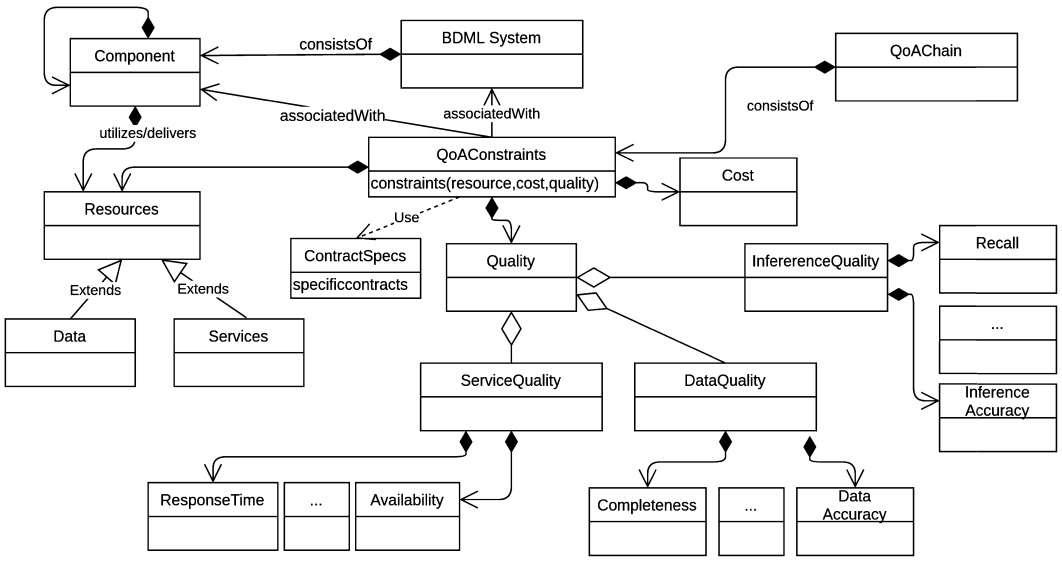


FIGURE 10.4 A simplified view of QoAChain and its relations to *BDML* systems and existing contracts.

tem. Therefore, a QoA-based approach can help to deal with the diversity of what, when, where, and how runtime attributes related to R3E can be supported. The QoA-based approach should include metrics for services, data, and ML models to reflect the end-to-end view. It should support human-in-the-loop and domain expert integration when defining QoA, due to the domain aspect of end-to-end *BDML*. To this end, we define "Chaining QoA for *BDML*" (QoAChain) as a contractual means for optimizing end-to-end *BDML* systems. QoAChain constraints described in a "contract" for optimizing R3E attributes (i) implement service contract and data contract models, (ii) enable monitoring and optimization techniques centered around contracts, and (iii) allow runtime changes and updates according ML-specific contexts by people or intelligent software. QoAChain constraints are based on various metrics inherent in *BDML*.

Fig. 10.4 shows key sub-elements of a proposed `QoAChain` and its relation to a *BDML* system. First, in our view, a `BDML System` consists of many `Components`; each `Component` may have sub components. A `Component` will utilize some resources (to implement required functions) and/or will deliver resources (e.g., featuring data/resulting prediction). `Resources` in our view can

be simple or complex, and they are not just infrastructural resources. The main categories of resources to be utilized or delivered are `Services` and `Data`. `Services` can be further divided into different types, such as for data processing, computing, storage, and inferencing. `Data` can be used to represent input data and output data (e.g., inference result in the case of ML service). In terms of quality, represented by `Quality`, there are many attributes known in big data and ML, such as `ResponseTime`, `Data Quality`, and `Inference Quality` (see also Section 10.2.1); we just illustrate some of them in Fig. 10.4. `QoAChain` for a `BDML System` consists of different `QoAConstraints`, which are associated with `Components` or the `BDML System` as a whole. `QoAConstraints` are used to specify constraints on attributes that should be monitored and optimized for R3E. They include tradeoffs among `Resources`, `Quality`, and `Costs`. `QoAConstraints` can be implemented by using existing, specific contract specifications. Examples of constraints in a chain are:

- a constraint on data completeness for IoT input data sent to a message broker, which passes the data to an ML service for dynamic inference of the IoT data
- a constraint on inference accuracy for an ML service, given a constraint on data completeness and data volume that the ML service handles in a window of time
- a constraint on the response time between from the time a component sends a batch of data to a message broker to an ML service until the time the component receives the inference result

Based on QoAChain, the next question is how to manage R3E attributes across multiple contexts in end-to-end *BDML* systems, such as to which components we should associate QoAChain and how to manage them.

10.3.2 R3E objects and operations

For **RQ2**, we will address fundamental abstractions for objects and operations for R3E. Consider the internal structure of a *BDML*: $BDML = \{c_1, c_2, \cdots, c_n\}$, whereas c_i is a component, which is a part of *BDML*. A component can be a software service, a container instance, a virtual machine (VM), or a middleware; a component can be instantiated as a resource-

as-a-service. A component can be composed from a set of components, creating a complex component as a subsystem of a *BDML* system. For example, a subsystem for data preprocessing in a *BDML* can include containers and workflow orchestration components. Given the structure of *BDML* explained in Section 10.2.1, a component of *BDML* can be described using a set of objects; an object can represent a very complex component, such as an extract-transform-load process that filters data suitable for feature engineering, or represent a simple task, e.g., a data validation task.

10.3.2.1 Conceptualize R3E objects

In terms of management, we view components, pipelines, tasks and their input/output as programmable *objects*. We define an object as *an R3E object* if we can associate R3E policies and attributes with the object, meaning that we can examine R3E capabilities for the object and control these attributes. Given a *BDML*, not all the objects can be an R3E object. Furthermore, in the view of the developer, they might not see an object as an R3E object if they cannot apply R3E techniques. However, the operator of *BDML* might see that object as an R3E one. For example, consider an ML model which has no elastic parameters to influence robustness. The developer might not focus on the ML model as an R3E object. The operator sees that the underlying computing resources can be changed for the ML model, thus it can be an R3E object.

We propose to conceptualize R3E objects, shown in Fig. 10.5. An `R3Eobject` represents a `BDML Component`. `BDML Component` can be classified according to their functionality and layers, such as infrastructural objects, ML algorithm objects, and data objects. Furthermore, an `BDML Component` can be composed from other `BDML Components`. Therefore we have a similar classification of `R3Eobject`. An `R3EObject` will have to implement a set of operations/APIs for controlling and monitoring and will be associated with a set of attributes —`R3EAttributes`— each attribute is represented as a metric name and value.

When applying the R3E approach, R3E objects can be identified and built from two perspectives: existing knowledge about contemporary objects that we use, such as containers, VMs, and middleware, which already have

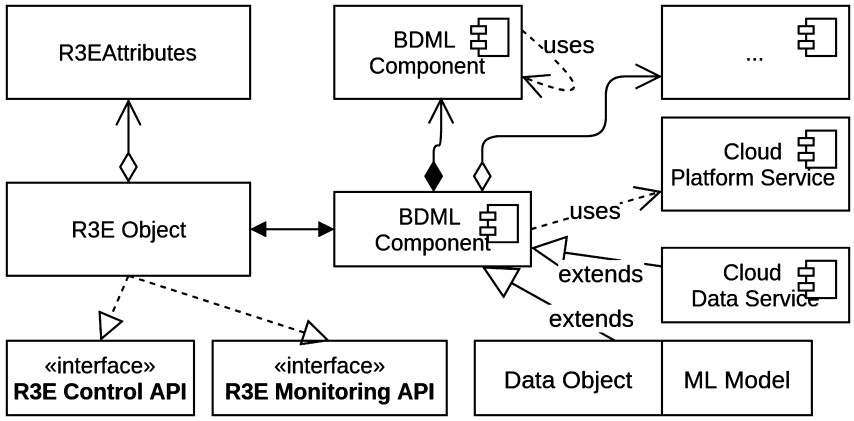


FIGURE 10.5 A simplified conceptual model of R3E objects.

built-in features for controlling certain aspects of R3E. Second, the new objects to be developed for *BDML* must implement such features. Given a *BDML* system, in our approach, we do not need to represent `R3EObject` for all possible `BDML Components`. However, using a composition model, we can also build an `R3EObject` for the entire *BDML* system that links to other `R3EObjects` representing other components. Via the dependency of `R3EObjects`, we can capture the whole picture of the system to be optimized. For the implementation, we are investigating two models: `R3EObject` as implemented as a resource of a microservice (an adaptor model) and as an interface implemented within components themselves.

10.3.2.2 R3E attributes associated with R3E objects

We classify attributes into different subcategories, associated with services, data, and ML models, shown in Fig. 10.4. Each attribute is represented as a metric under a tuple $(name, value)$.

- Services quality: covers different types of attributes for a variety of services, including infrastructural computing services, data storage, communication services, and platform services. Common quality attributes are well-known in literature, such as response time, availability, and MTBF.

- Data quality: covers data quality metrics, such as completeness, timeliness, currency, validity, format, accuracy, and data drift.
- ML models quality: includes known quality in ML models, such as accuracy, F1 Score, and MSE.

These metrics are captured for individual components and composite components and tasks of ML pipelines carried out atop such components.

10.3.2.3 R3E operations and APIs

Given an R3E object, we must be able to control it to meet R3E constraints, which are pre-defined or changed during runtime. For example, if parameters of an ML model as a R3E object can be controlled to affect the ML model, we can then optimize the ML model for different degrees of robustness, reliability, resilience, and elasticity. Similarly, if an object performing feature engineering can be tuned with different granularity of feature extraction and selection, then we can control the object to have different data quality values. Furthermore, to allow for controlling, we must be able to monitor and query states of R3E objects at runtime. This can be done directly through querying the object or indirectly through the monitoring systems. Shown in Fig. 10.5, two types of key operations are for R3E controlling and monitoring. An R3E operation associated with R3E objects will be implemented as an API. Inputs and outputs of the API are centered around metrics and constraints specified in QoAChain.

10.3.3 Engineering methods

For **RQ3**, we propose a set of engineering methods for R3E coordination, monitoring and analytics, and testing, benchmarking and experiments. We will describe engineering methods, but leave the implementation of tools and frameworks for such methods out of the scope of this chapter.

10.3.3.1 Coordination for R3E

Having R3E objects enables us to optimize the *BDML* system through control and reconfiguration of R3E attributes, thus leading to changes in components of the *BDML* system. Due to the complexity and structure of the

BDML system, coordination of such controls and configurations is challenging.

Architectural styles for R3E coordination: To perform the coordination of controls and reconfigurations of various R3E objects, we must consider suitable architectural styles coupled with *BDML* systems. Most architectures for end-to-end *BDML* systems follow either the reactive style or the workflow style as the basic architectural style. Furthermore, due to the complexity of individual components, each component might also follow the workflow or reactive style. Basically,

- reactive style: the data/event from one task/component triggers the next action in the pipeline/system (Smith, 2018). This model usually fits very well with large-scale *BDML* systems
- workflow style: a workflow is used to control tasks/components in ML pipelines/systems. However, most systems focus on leveraging workflows in the training or inferencing.

We design our approach to work with the reactive system style. This will be aligned with *BDML* pipelines consisting of components across different layers and different infrastructures (e.g., edge and cloud) and different providers. Furthermore, using reactive models, we can intercept *BDML* systems at different places to create optimization and feedback channels to support R3E. Fig. 10.6 present the high-level components view. A *BDML* system consists of different subsystems, such as *Computing and Data Platform*, *Data Processing Platform*, and *Serving Platform*. Each subsystem is complex and can be implemented with different technologies. ML tasks in ML pipelines are spread in these subsystems, and they are coupled through reactive principles by using messages. Therefore we do not need a global workflow system to orchestrate them, but we can use a set of `R3E Managements`. Each `R3E Management` will interact with a subsystem using three interfaces: `R3EPolicies` are used to control subsystems; `R3EAttributes` are used to capture states, and `R3EConstraints` specifies QoAChain. Among `R3EManagements`, the reactive principle is also used to provide an end-to-end view of the whole system.

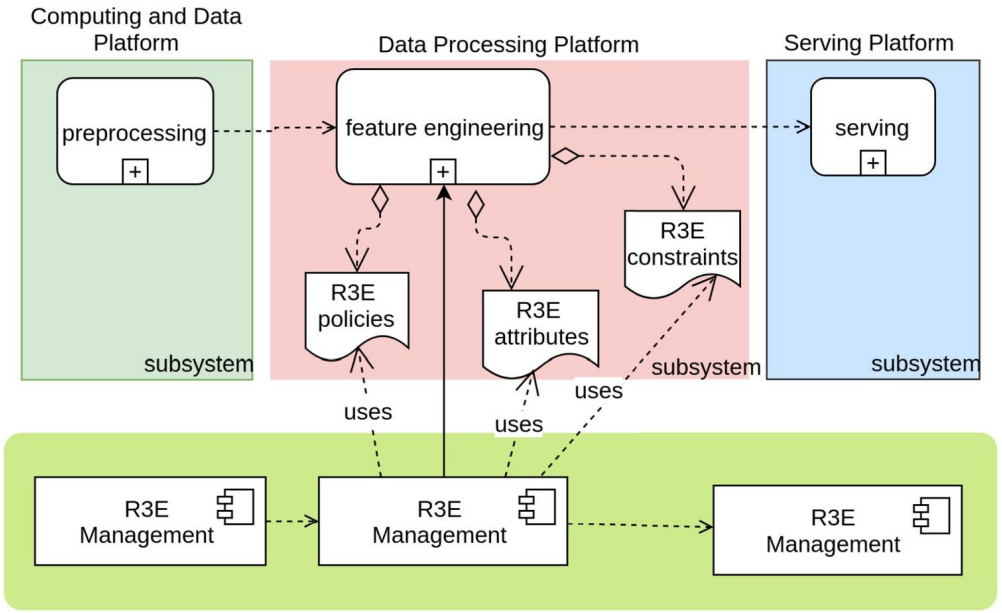


FIGURE 10.6 R3E reactive systems with group of management.

Distributed controls: Each component, based on the view in Fig. 10.1, can be controlled and managed through the individual component's R3E object. For example, data collections can be controlled to select suitable data sources, and such controls are independent from another control of the ML model service. However, from an end-to-end viewpoint, we need to coordinate these controls to achieve the defined QoAChain for the whole *BDML* system. Fig. 10.7 presents our approach to control ML tasks and corresponding components by using `Monitoring Process` and `Control Process` to interact with `R3EObject`. For each subsystem and subpipelines on that subsystem, `R3EObjects` are used to monitor and control tasks and components. Corresponding `Monitoring Process` and `Control Process` will interact with these `R3Object`. Exchanges among subpipelines, such as the featuring data outputted from `feature engineering` subpipeline to `training` subpipeline, will be also monitored. Individual `Control Process` for different subpipelines will be coordinated by a crossover `Control Process`. Thus we have distributed controls, but through a centralized coordination. In our implementation, we plan to implement functions in `Monitoring Pro-`

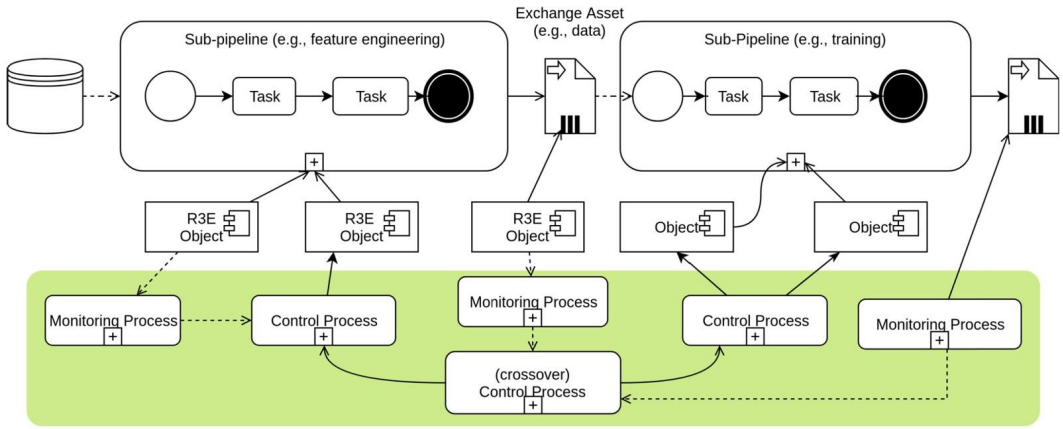


FIGURE 10.7 Distributed monitoring and control processes utilizing R3E objects.

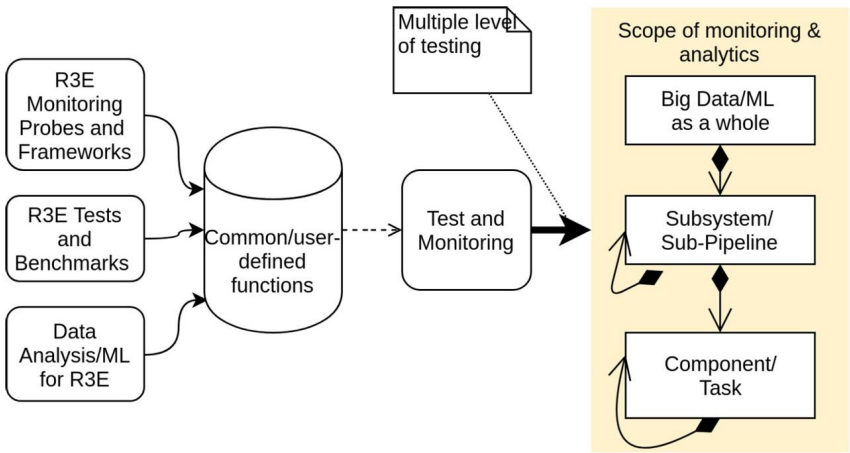


FIGURE 10.8 Scopes of tests, benchmarks and experiments.

`cess` and `Control Process` using serverless frameworks. Note that `Monitoring Process` will need to work with monitoring systems that we will discuss in the next section.

10.3.3.2 Monitoring and analytics

Monitoring and analytics monitor and analyze R3E attributes of services, data, and ML models and map the R3E attributes to R3E objects. For the whole approach, we need to leverage different methods for monitoring and analytics. Shown in Fig. 10.8, we will need (i) monitoring probes and frame-

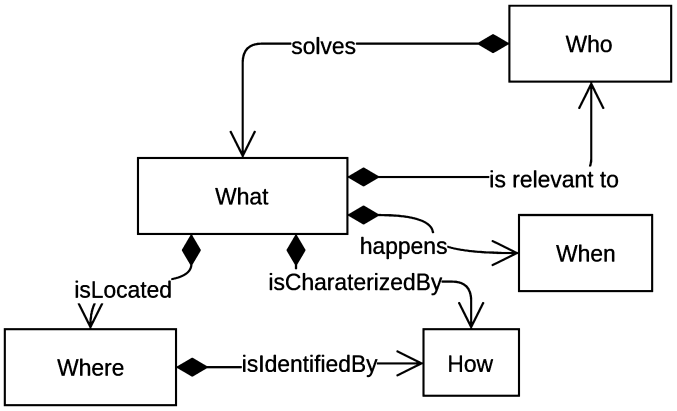


FIGURE 10.9 R3E W4H analytics context.

works, (ii) tests and benchmarks units and frameworks, and (iii) data analysis/machine learning for understanding monitoring data. We will support different scopes of monitoring and analytics: the system as a whole, subsystem/subpipeline, and component/task. With such scopes, we will focus on end-to-end aspects. For example, data reliability can be examined along the path from data sources to the final inference results. The consumer can also expect an end-to-end R3E attribute, such as accuracy and response time, which can only be achieved if we are able to monitor different parts and to perform coordination-aware assurance, e.g., using elasticity principles, in the system as a whole scope. For the implementation, we will need to integrate various monitoring systems for services (such as Prometheus), for data (e.g., data validation tools from `scikit-learn` and TensorFlow Data Validation), and for ML models (e.g., extracted from ML frameworks).

In terms of analytics, we also have different perspectives about which techniques can be applied for which parts and whether we can have evaluation and interpretation in a subjective manner. Here the context of analytics is important. Therefore we suggest to use a context model of what, when, where, who and how for analyzing R3E. Fig. 10.9 presents the context model for understanding R3E attributes. Our approach in the implementation is to extend the work on monitoring of ML service contracts (Truong and Nguyen, 2021) to cover R3E.

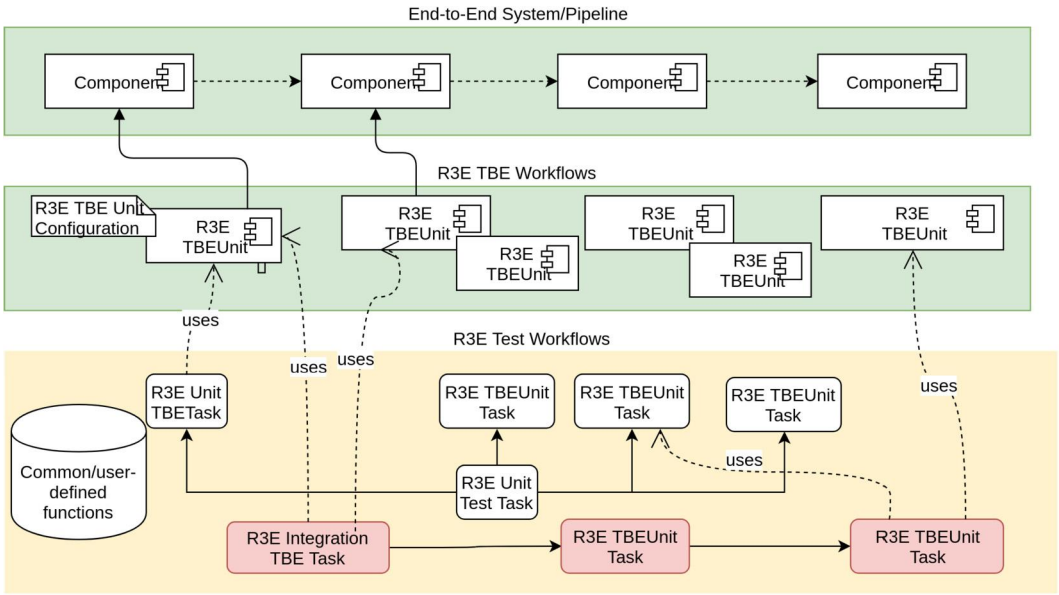


FIGURE 10.10 Conceptual view of R3E tests, benchmarks, and experiments.

10.3.3.3 Testing, benchmarking, and experimenting for R3E

To run tests, benchmarks, and experiments (TBE) is of paramount importance for optimizing R3E. The challenge is that it is not just related to ML training and serving, but also to other tasks in the whole *BDML* system. Testing, benchmarking, and experimenting have to do across subsystems and focus on correlating R3E issues. Current ML testing frameworks are mainly focused on ML models (Aggarwal et al., 2019; Riccio et al., 2020). Testing big data is currently focused on data storage and querying (Alexandrov et al., 2013; Li et al., 2016; Baru et al., 2012; Gulzar et al., 2019; Bajaber et al., 2020). Combination of different tests into a coherent R3E view is currently missing. Furthermore, ML experiment solutions (Gharibi et al., 2019; Duarte et al., 2017) focus on mainly model metadata, used datasets, and hyperparameters, but the management of data sources, services performances, and code/data versions is not integrated.

Fig. 10.10 outlines our approach for testing, benchmarking, and experimenting. `R3E TBEUnit` is an abstract unit designed for testing, benchmarking, and experimenting. At the top level, we use workflows to coor-

dinate tests/benchmarks across subsystems for different subpipelines. We will develop a variety of R3E test/benchmark units for *BDML*; each unit test/benchmark not only one component, but also a layer or an aspect, e.g., data. Last, R3E integration tests/benchmarks/experiments carried out for individual subsystems and components are linked together.

10.4 Illustrative examples

In this section, we explain strategies for optimizing the BIM scenario mentioned in Section 10.2.2, considering that we have to apply the R3E approach for the BIM scenario to allow the optimization of the BIM ML pipeline in an end-to-end manner. To this end, we analyze the scenario and apply step-by-step of the R3E approach mentioned in Section 10.3. In what follows, we summarize R3E aspects, identified requirements to be addressed. For each category, we only show key examples.

In terms of **QoAChain** we have identified:

Aspects	**Identified requirements**
Model accuracy	as a contractual means between BIM and customers for the whole system
Response time	as a contractual means between the ML service provider and customers for inferencing
Accuracy & response time tradeoffs	accuracy is more important than response time to customers. Therefore elasticity of cloud resources for ML services can be flexible (using CPU, GPU, and even spot instances)

The accuracy and response time tradeoffs are based on the business of the BIM scenario: it is important to predict and classify building objects with a high degree of accuracy to make sure that the design is correct and reliable. For this, the customer does not need a real-time prediction. Consequently, in terms of R3E at runtime, computing and data resources could be allocated differently with respect to the cost. For example, GPU resources might be used only if a higher cost is accepted by the customer, whereas more computing resources are needed for data preprocessing and feature engineering phases.

In terms of **R3E objects and operators**, we have identified:

Aspects	Identified requirements
R3E objects	include (i) data resources collector and selector (trustful data sources, text, and 3D data); (ii) feature engineering component (3D data extraction granularity); (iii) ML models (model versions and parameters); (iv) ML services (coupled with ML models and underlying computing resources); computing resources (cloud-based CPU & GPU resources)
Operators	data feature engineering operators are for fine-grained and high-grained of data extraction; changes of ML models and parameters; elastic computing resources with possibility to have different types of resources, including edge hardware, cloud-based CPU and GPU

The data resource collector and selector not designed as data sources are typically fixed, e.g., from S3 storage or shared file-based storage. This requires the design of new components for data collector and selector that are integrated with data source metadata and data resource catalogs and the change of the data pipeline from `User Tool` to `Data Service` (see Fig. 10.2). One example is to use DVC[1] in combination with quality of data metrics, such as trust, data completeness, and timeliness, that are determined during the data export task. Another aspect is to control `feature engineering` task, which is tightly coupled with `preprocessing`, but strongly influences the accuracy, cost, and time of the prediction in ML models. This requires us to separate *preprocessing* and *feature engineering* tasks.

In terms of **R3E Engineering**, we have identified:

Aspects	Identified requirements
Coordination for R3E	three subsystems are identified: preprocessing, featuring engineering, and serving. They can be optimized independently or together.
Monitoring and analytics	quality of data in data collection; model accuracy metrics during serving; performance response time; costs paid to cloud resources
Tests, benchmarks, and experiments	tests of data validation; benchmarks in training; monitoring and provenance for data resources and machines for individual experiments

Through the separation between *preprocessing* and *feature engineering*, and introduce an R3E object between the two tasks for controlling `feature engineering`. Finally, ML models need to be controlled separately through QoAChain coupling the ML models with resource elasticity.

[1] https://dvc.org.

10.5 Discussion

Different communities have advocated R3E in different ways. In ML benchmarks, various initiatives for testing robustness, performance, and cost have been carried out (Tang et al., 2020). Recently the discussion of end-to-end ML has attracted many researchers. Especially, in the software engineering and ML production, the optimization of many phases in end-to-end ML pipelines is on the focus (Amershi et al., 2019). AIOps (Dang et al., 2019) focuses on using AI to optimize quality of software. Our work has a similar ultimate goal, but our approach differs; we focus on software service programming, engineering, and analytics. To our best knowledge, there is no previous roadmap for R3E for end-to-end *BDML*.

Different components of an end-to-end *BDML* system have different R3E attributes and concerns that lead to different optimization focuses of R3E. For example, in ML services and models, robustness as a critical concern has been discussed intensively (Sehwag et al., 2019). However, the elasticity of ML services has not been studied well (Huang et al., 2015), while the elasticity of infrastructural resources for enabling cloud computing has been studied intensively. Another example is that, while data is very important for robustness in ML training and ML services, monitoring quality of data monitoring and supporting data resources elasticity in ML have not been well developed. The reliability, reflecting the concept of offering "reliable service" (Kumar and Vidhyalakshmi, 2018; Galetzka et al., 2006), for different individual components (e.g., ML services, data stores, and data brokers) has been studied intensively, but the reliability of *BDML* systems has not been studied well in an end-to-end manner. Resilience (Robbins et al., 2012) is mostly addressed at the system services. We need to understand how ML models resilience (Park et al., 2017) is related to elasticity of data and other aspects, e.g., message middleware (Wang et al., 2010) and programming languages (Grove et al., 2019) in our pipeline design.

10.6 Conclusions and future work

In this chapter, we present a novel conceptual approach for implementing robustness, reliability, resilience, and elasticity for end-to-end big data

machine learning systems, called R3E. Given R3E attributes, we proposed to use QoAChain as a contractual means for specifying constraints of R3E. To manage R3E of components and tasks, from the R3E view, we abstract them under R3E objects and devise operations for monitoring, controlling, and optimizing R3E. Our approach has presented key engineering methods for main design and engineering activities with respect to R3E. In our current work, we have not addressed all tools that implemented our abstractions and engineering methods. We are working on the monitoring and observability of ML services (Truong and Nguyen, 2021) and a service for collecting trails from tests, benchmarks, and experiments for end-to-end ML systems. However, this chapter layouts fundamental steps for addressing R3E design and engineering in the future.

We foresee that different scenarios can be elaborated to have a deeper view on R3E. Details of tools and components can be carried out for training optimization, runtime ML model serving, out-of-distribution detection and optimization, and elasticity serving. The situation is even more challenging when *BDML* systems have more distributed learning (Verbraeken et al., 2020), as the nature of complexity is increasing. We will revise our approach to address such new development. Our current work is to focus on two aspects: end-to-end self-optimized solutions and QoAChain toolset.

Acknowledgments

We are grateful to Minjung Ryu for the discussion about the case of ML for classification of building information modeling (BIM) elements.

References

Ackley, D.H., 2013. Beyond efficiency. Communications of the ACM 56, 38–40. https://doi.org/10.1145/2505340.

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D., 2019. Black box fairness testing of machine learning models. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. Association for Computing Machinery, New York, NY, USA, pp. 625–635.

Alexandrov, A., Brücke, C., Markl, V., 2013. Issues in big data testing and benchmarking. In: Proceedings of the Sixth International Workshop on Testing Database Systems. Association for Computing Machinery, New York, NY, USA.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T., 2019. Software engineering for machine learning: a case study. In: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice. IEEE Press, pp. 291–300.

Bajaber, F., Sakr, S., Batarfi, O., Altalhi, A., Barnawi, A., 2020. Benchmarking big data systems: a survey. Computer Communications 149, 241–251.

Balint, F., Truong, H.L., 2017. On supporting contract-aware iot dataspace services. In: 5th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering. MobileCloud 2017, San Francisco, CA, USA, April 6–8, 2017. IEEE Computer Society, pp. 117–124.

Baru, C., Bhandarkar, M., Nambiar, R., Poess, M., Rabl, T., 2012. Big data benchmarking. In: Proceedings of the 2012 Workshop on Management of Big Data Systems. Association for Computing Machinery, New York, NY, USA, pp. 39–40.

Blanchard, P., Mhamdi, E.M.E., Guerraoui, R., Stainer, J., 2017. Machine learning with adversaries: byzantine tolerant gradient descent. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 4–9 December 2017, Long Beach, CA, USA, pp. 119–129.

Brtis, J., Jackson, S., Cureton, K., 2021. https://www.sebokwiki.org/wiki/System_Resilience. (Accessed 15 October 2021).

Dang, Y., Lin, Q., Huang, P., 2019. Aiops: real-world challenges and research innovations. In: Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings. IEEE Press, pp. 4–5.

Duarte, J.C., Cavalcanti, M.C.R., de Souza Costa, I., Esteves, D., 2017. An interoperable service for the provenance of machine learning experiments. In: Proceedings of the International Conference on Web Intelligence. Association for Computing Machinery, New York, NY, USA, pp. 132–138.

Dustdar, S., Guo, Y., Satzger, B., Truong, H.L., 2011. Principles of elastic processes. IEEE Internet Computing 15, 66–71.

Elsayed, E.A., 2012. Reliability Engineering, 2nd ed. Wiley Publishing.

Fischer, L., Memmen, J., Veith, E.M.S.P., Tröschel, M., 2018. Adversarial resilience learning - towards systemic vulnerability analysis for large and complex systems. CoRR. arXiv:1811.06447. arXiv:1811.06447.

Galetzka, M., Verhoeven, J., Pruyn, A., 2006. Service validity and service reliability of search, experience and credence services: a scenario study. International Journal of Service Industry Management 17, 271–283.

Gharibi, G., Walunj, V., Rella, S., Lee, Y., 2019. Modelkb: towards automated management of the modeling lifecycle in deep learning. In: Proceedings of the 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering. IEEE Press, pp. 28–34.

Gribble, S.D., 2001. Robustness in complex systems. In: Proceedings Eighth Workshop on Hot Topics in Operating Systems, pp. 21–26.

Grove, D., Hamouda, S.S., Herta, B., Iyengar, A., Kawachiya, K., Milthorpe, J., Saraswat, V., Shinnar, A., Takeuchi, M., Tardieu, O., 2019. Failure recovery in resilient x10. ACM Transactions on Programming Languages and Systems 41.

Gujarati, A., Elnikety, S., He, Y., McKinley, K.S., Brandenburg, B.B., 2017. Swayam: distributed autoscaling to meet slas of machine learning inference services with resource efficiency. In: Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference. Association for Computing Machinery, New York, NY, USA, pp. 109–120.

Gulzar, M.A., Mardani, S., Musuvathi, M., Kim, M., 2019. White-box testing of big data analytics with complex user-defined functions. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. Association for Computing Machinery, New York, NY, USA, pp. 290–301.

Harlap, A., Tumanov, A., Chung, A., Ganger, G.R., Gibbons, P.B., 2017. Proteus: Agile ml elasticity through tiered reliability in dynamic resource markets. In: Proceedings of the Twelfth European Conference on Computer Systems. Association for Computing Machinery, New York, NY, USA, pp. 589–604.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2017. Deep reinforcement learning that matters. In: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2018.

Hendrycks, D., Dietterich, T.G., 2019. Benchmarking neural network robustness to common corruptions and perturbations. CoRR. arXiv:1903.12261. arXiv:1903.12261.

Huang, B., Boehm, M., Tian, Y., Reinwald, B., Tatikonda, S., Reiss, F.R., 2015. Resource elasticity for large-scale machine learning. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery, New York, NY, USA, pp. 137–152.

Katzir, Z., Elovici, Y., 2018. Quantifying the resilience of machine learning classifiers used for cyber security. Expert Systems with Applications 92, 419–429.

Kleppmann, M., 2016. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems. O'Reilly.

Kulkarni, A., Chong, D., Batarseh, F.A., 2020. 5 - foundations of data imbalance and solutions for a data democracy. In: Batarseh, F.A., Yang, R. (Eds.), Data Democracy. Academic Press, pp. 83–106. https://www.sciencedirect.com/science/article/pii/B9780128183663000058.

Kumar, V., Vidhyalakshmi, R., 2018. Reliability Aspect of Cloud Computing Environment, 1st ed. Springer Publishing Company, Incorporated.

Laranjeiro, N., Agnelo, J.a., Bernardino, J., 2021. A systematic review on software robustness assessment. ACM Computing Surveys 54. https://doi.org/10.1145/3448977.

Lee, D., 2019. Big data quality assurance through data traceability: a case study of the national standard reference data program of Korea. IEEE Access 7, 36294–36299.

Li, N., Lei, Y., Khan, H.R., Liu, J., Guo, Y., 2016. Applying combinatorial test data generation to big data applications. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering. Association for Computing Machinery, New York, NY, USA, pp. 637–647.

Littlewood, B., Strigini, L., 2000. Software reliability and dependability: a roadmap. In: Proceedings of the Conference on the Future of Software Engineering. Association for Computing Machinery, New York, NY, USA, pp. 175–188.

Nguyen, C., Mehta, A., Klein, C., Elmroth, E., 2019. Why cloud applications are not ready for the edge (yet). In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. Association for Computing Machinery, New York, NY, USA, pp. 250–263.

Park, S., Weimer, J., Lee, I., 2017. Resilient linear classification: an approach to deal with attacks on training data. In: Proceedings of the 8th International Conference on Cyber-Physical Systems. Association for Computing Machinery, New York, NY, USA, pp. 155–164.

Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., Tonella, P., 2020. Testing machine learning based systems: a systematic mapping. Empirical Software Engineering 25, 5193–5254. https://doi.org/10.1007/s10664-020-09881-0.

Robbins, J., Krishnan, K., Allspaw, J., Limoncelli, T., 2012. Resilience engineering: learning to embrace failure. Communications of the ACM 55, 40–47. https://doi.org/10.1145/2366316.2366331.

Ryu, M., Truong, H.L., Kannala, M., 2021. Understanding quality of analytics trade-offs in an end-to-end machine learning-based classification system for building information modeling. Journal of Big Data 8, 31. https://doi.org/10.1186/s40537-021-00417-x.

Saria, S., Subbaswamy, A., 2019. Tutorial: safe and reliable machine learning. CoRR. arXiv:1904.07204. arXiv:1904.07204.

Sehwag, V., Bhagoji, A.N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., Mittal, P., 2019. Analyzing the robustness of open-world machine learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. Association for Computing Machinery, New York, NY, USA, pp. 105–116.

Smith, J., 2018. Machine Learning Systems: Designs That Scale, 1st ed. Manning Publications Co., USA.

Suryavansh, S., Bothra, C., Chiang, M., Peng, C., Bagchi, S., 2019. Tango of edge and cloud execution for reliability. In: Proceedings of the 4th Workshop on Middleware for Edge Clouds & Cloudlets. Association for Computing Machinery, New York, NY, USA, pp. 10–15.

Tang, Fei, et al., 2020. AIBench: an Industry Standard AI Benchmark Suite from Internet Services. Technical Report. BenchCouncil: International Open Benchmarking Council.

Trivedi, K.S., Kim, D.S., Ghosh, R., 2009. Resilience in computer systems and networks. In: 2009 IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers, pp. 74–77.

Truong, H.L., Comerio, M., Paoli, F.D., Gangadharan, G.R., Dustdar, S., 2012. Data contracts for cloud-based data marketplaces. International Journal of Computational Science and Engineering 7, 280–295. https://doi.org/10.1504/IJCSE.2012.049749.

Truong, H.L., Murguzur, A., Yang, E., 2018. Challenges in enabling quality of analytics in the cloud. Journal of Data and Information Quality 9. https://doi.org/10.1145/3138806.

Truong, H.L., Nguyen, T., 2021. Qoa4ml – a framework for supporting contracts in machine learning services. In: 2021 IEEE International Conference on Web Services (ICWS). IEEE, United States.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S., 2020. A survey on distributed machine learning. ACM Computing Surveys 53.

Wang, J., Balazinska, M., 2017. Elastic memory management for cloud data analytics. In: Silva, D.D., Ford, B. (Eds.), 2017 USENIX Annual Technical Conference. USENIX ATC 2017, Santa Clara, CA, USA, July 12–14, 2017. USENIX Association, pp. 745–758.

Wang, J., Jiang, P., Bigham, J., Chew, B., Novkovic, M., Dattani, I., 2010. Adding resilience to message oriented middleware. In: Proceedings of the 2nd International Workshop on Software Engineering for Resilient Systems. Association for Computing Machinery, New York, NY, USA, pp. 89–94.

Yang, F., Chien, A.A., Gunawi, H.S., 2017. Resilient cloud in dynamic resource environments. In: Proceedings of the 2017 Symposium on Cloud Computing. Association for Computing Machinery, New York, NY, USA, p. 627.

Zhang, A., Song, S., Wang, J., Yu, P.S., 2017. Time series data cleaning: from anomaly detection to anomaly repairing. Proceedings of the VLDB Endowment 10, 1046–1057.

Zhang, P., Cao, W., Muccini, H., 2020. Quality assurance technologies of big data applications: a systematic literature review. CoRR. arXiv:2002.01759. arXiv:2002.01759.

Zheng, Y., Xu, L., Wang, W., Zhou, W., Ding, Y., 2017. A reliability benchmark for big data systems on jointcloud. In: 2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 306–310.

This page intentionally left blank

PART 3

AI assurance and applications

This page intentionally left blank

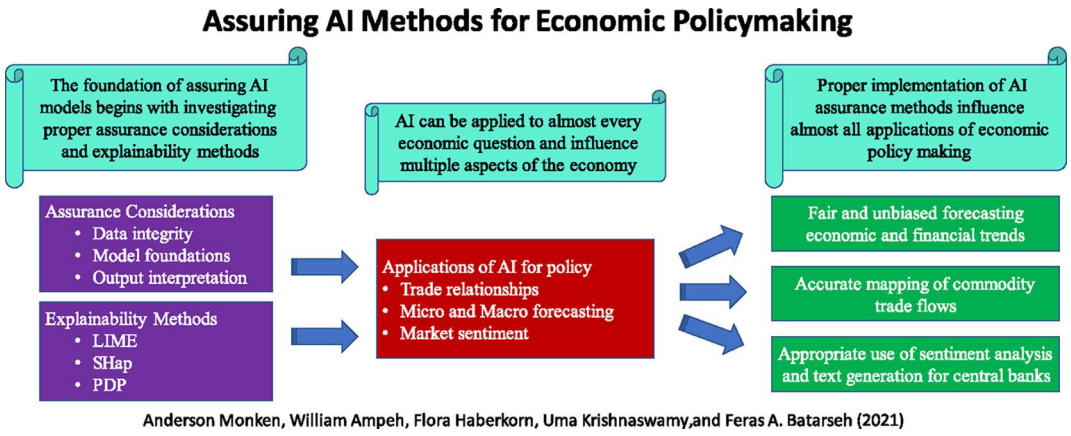
11

Assuring AI methods for economic policymaking

Anderson Monken[a,b,g], William Ampeh[c,d,g], Flora Haberkorn[a,g], Uma Krishnaswamy[a,e,g], and Feras A. Batarseh[f]

[a]International Finance Division of the Federal Reserve Board, Washington D.C., United States [b]Georgetown University, Washington D.C., United States [c]Research and Statistics Division of the Federal Reserve Board, Washington D.C., United States [d]George Mason University, Fairfax, VA, United States [e]University of California, Berkeley, Berkeley, CA, United States [f]Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States

Graphical abstract



[g]The views expressed in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System.

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00025-1

Abstract

AI methods are becoming more common in the field of economics, but these models must be bias-free, fair, and explainable. In other words, we need AI assurance. Economic forecasting has benefited from machine learning techniques, such as neural networks, to increase model performance, but these AI techniques must be audited, accountable, and interpretable to be useful for economic policymaking. The rise of natural language processing and large language models has created new challenges for economic policymaking institutions, which need to be aware of AI assurance and how to harness them safely.

Keywords

AI assurance, AI explainability, detecting bias, natural language processing, large language models

Highlights

- Discover how Artificial Intelligence is being used in economics and how the field can mitigate bias by following assurance practices in both sectors, private, and public
- Get exposed to an in-depth overview of explainability techniques in microeconomic research with sample code and graphical output
- Find out how natural language processing is implemented in economics and mitigate potential risks for typical tasks such as conducting sentiment analysis and utilizing large language models for text generation

11.1 Introduction to harnessing AI for economics

Economics and Artificial Intelligence (AI) are becoming increasingly linked as increased computational power and advanced AI models are supplementing the traditional realms of econometrics. To fully embrace the potential of AI in economics, the "black box" needs to be cracked open to assure that more flexible machine learning (ML) models can be explained as effectively as traditional linear models.

Economic policymaking can be enhanced with the use of AI. It is an exciting alternative to econometric equilibrium models with unrealistic model agents. AI can solve nonlinear and nuanced problems in economics that are out of reach to traditional econometrics. AI in economic simulations can

act like human actors, evading policies for individual benefit similar to real behavior. In Zheng et al. (2020), improvements in tax policy are evaluated using reinforcement learning agents who learn how to stockpile income to avoid taxation and choose their own level of output given tax rates. The AI tax economist developed tax rates to improve both equality and productivity in their simulated world compared to real-world outcomes. What is the potential for economics if AI continues to advance at its current pace? What are the risks to relying on these models in research and policy? How can AI be used in economic institutions?

AI produced models for economic policymaking need assurance through the entire process: ethical data collection and pre-processing, bias mitigation during and after model training, and explainability techniques for model outcomes. AI has helped advance the study of economics, while economic theories, particularly in the field of game theory, have also improved AI explainability through tools such as SHapley Additive exPlanations (SHAP) (Shapley, 1953; Lundberg and Lee, 2017).

This chapter examines the landscape of AI use in economic modeling, natural language processing (NLP) applications for economics, the possibilities for gain and misuse of large language models and its impact on policymaking, as well as discussing the innovations of AI for international trade. Policymaking transparency, applications of explainability methods, and concerns about model assurance will be a central theme throughout these topics.

AI research has gained attention by designing simulation models to forecast stock fluctuations and other microeconomic market measures. Parkes and Wellman (2015) reviews how AI researchers are working towards a "synthetic homo economicus," an economic agent capable of making rational decisions, along with the challenges and progress made in the design of these intelligent agents. As economic researchers continue to explore ways to improve the performance of standard econometric models, their attention continues to be drawn to the use of ML algorithms as these methods continue to see dramatic improvements in information technology and other research fields.

The impact of AI in economic theories and the current technology holders are documented by Marwala and Hurwitz (2017) and Moloi and Marwala (2020). As with other fields of study, the availability of ever larger datasets has resulted in the increasing use of AI techniques, such as machine learning in the fields of microeconomics and finance.

The availability of massive datasets (big data) have proved to be essential for sound policy decision-making in both the public and private sectors, with many private organizations viewing big data as not just a buzzword, but instead as a long-term strategic concept to build into their workflow. The potential benefits of big data, from macroeconomic and financial statistics to ultimately policymaking, are documented by Hammer et al. (2017). Khalid and Rachid (2019) discuss uses of big data in economic analysis and their potential to improve microeconomic prediction models accuracy, while also identifying the challenges of using big data in economic analysis, which include access, replication, and the technological skill to work with these datasets.

Big data analytics are being eagerly explored by economists, and the use of AI techniques on big data is considered one of the greatest potentials for new discoveries in economic analysis. The accessibility of ML methods in trending programming software languages, such as R and Python, and the ability to link these data analytics to other conventional programming languages, such as SAS and MATLAB®, continue to increase the popularity of ML for econometric analysis. For example, one can easily download software (for either R or Python) and, with a few lines of code, fit decision trees or random forests regression on a large dataset and quickly extract the relevant coefficients.

Natural language processing (NLP) techniques have permeated nearly every industry and have successfully made their way into economic policy. NLP provides a rigorous way to create additional quantitative data to study the drivers of economic outcomes and changes to relevant trends. In this section, the need for transparent and methodical handling of textual analysis is highlighted, specifically in the context of assessing the effects of Federal Open Market Committee (FOMC) communications on financial asset prices and firms' earnings calls, trends in the wake of the COVID-19

global pandemic. There will also be discussions on how to assure black-box textual models that are so frequently employed on central bank statements, social media texts, and other large corpora over time. In addition, prior to the analysis stage of the AI lifecycle, thorough assurance and bias mitigation is required at the data collection, model training, and evaluation stages.

Many companies have utilized complex neural network techniques to create products and services to sell to consumers as well as help business operations. Language models are integrated into many tools, such as customer service chat boxes, language translation programs, and search engines. A well-known class of products that use language models are AI assistants that execute tasks based on user verbal commands, such as Apple's Siri and Amazon's Alexa programs. The leading drivers of language model development are the creation of tools for assisting persons with disabilities, overcoming language barriers, reducing time on tedious tasks, and developing intelligence in robotics. Over the last decade, innovation in AI and increased public use of neural networks has spread this technology across society. Thus making it very likely the average person has already encountered or used a language model whether they realized it or not.

Widespread use of language models in the field of economics and finance is still in its early stages. This is specifically evident in macroeconomics, where there are not many public mentions of how institutions are utilizing language models for policy analysis in their journey to monitor the aggregate economy. However, in the private sphere, language models are being tested to conduct sentiment analysis on monetary policy reports and to generate simulated Federal Reserve statements. There is great interest and incentive for using natural language processing (NLP) techniques on monetary policy announcements to detect "hawkish" sentiment that would imply central bankers being focused on policy measures surrounding inflation or "dovish" sentiment, which focuses on the labor market and economic growth. It is important to note that language models are complex and using them effectively requires heavy investments in education and application. These investments have a significant influence on a team's capabilities to utilize them within financial institutions and private banks. At this time, institutional economists and Wall Street data scientists often use

simpler NLP methods to answer textual related questions, but this could change in the near future, thanks to increasingly powerful language models.

Language models at their core can be defined as a statistical probabilistic model that determines the probability of a given sequence of words within a sentence. They are often used within NLP for executing tasks such as text summarization and translation. Large language models (LLMs) are a class of language models that use deep-learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size. LLMs can be classed into two types: generative or discriminatory. Generative LLMs are models that output text, such as the answer to a question or even writing an essay on a specific topic. They are typically unsupervised or semi-supervised learning models that predict what the response is for a given task. Discriminatory LLMs are supervised learning models that usually focus on classifying text, such as determining whether a text was made by a human or AI.

The use of Artificial Intelligence in the domain of international trade has grown as more access to advanced computing techniques and micro-level data become available. International trade has several scopes from macro-level trade balances at the country level to commodity-level data of bilateral country trade, to individual container level shipments traveling from one country to another, and to the GPS data on the ships tracking the routes of those containers. It is critical to have assurance for AI methods at the data, model structure, and outcome stages can effectively assess the value of these rich data sources.

The success of international trade has important implications for the global economy. Supply chains have become networks across national borders, with "just-in-time" manufacturing providing low inventory costs and high productivity. Even minor disruptions to trade can add substantial costs for consumers, producers, and shippers. As world borders shut down at the beginning of the COVID-19 pandemic, trade ground to a halt. This affected not only shipments of furniture and exercise bikes, but also personal protective equipment and pharmaceuticals. Data at the commodity level can show the distribution of this catastrophic halt to trade. Further complicat-

ing trade in 2021 was the grounding of the container ship, *Ever Given*, in the Suez Canal, which blocked trade through one of the busiest trade routes in the world. Data companies, such as Marine Traffic provided data on exact ship locations, which enabled analysts to evaluate the magnitude of the shipping delays. To effectively prepare for the next black swan event that impedes trade, data scientists must know how to harness and assure AI methods using this complex data.

The rise of big data has advanced the potential for international trade research using AI methods. Traditional data sources, such as national trade balances, have delays in release that limit usefulness as input for quick policy responses. The coarseness of overall trade data also obscures the underlying mechanisms at the commodity level, which is another example of Simpson's paradox. Big data tools, such as Hadoop and Spark, are beneficial for providing the needed computational power to process hundreds of millions of rows of container level data.

A variety of AI methods are used to study international trade data, including data mining and graph neural networks. Each of these methods requires specific assurance techniques to implement reliably.

11.1.1 ML in economic models

Advances in AI techniques (including ML) have encouraged its widespread use, ushering in "4IR or Industry 4.0" (Schwab, 2016), the fourth industrial revolution, where the rapid accumulation of data models and input data is the new norm (Moloi and Marwala, 2020). The accessibility of ML tools has also increased the risk of improperly utilizing ML toolkits or misinterpreting their results. The increasing desire to improve the performance of ML models has led to the production of more complex models, which require additional tools to allow for reliable and clear explanations. These different tools can take textual and visual indicators that give users a realistic understanding of the relationship between the predictors and the model's generated prediction.

These explanation tools are considered critical in economic research, because many economic models depend on determining accurate parameters (such as $\beta_0, \beta_1, \ldots, \beta_n$) to predict the relationship between a response

variable (y) and a set of explanatory variables (X). However, many machine learning algorithms rely on producing predictions of y using complex and multi-layered parameters that are not clearly associated with a specific explanatory variable x_i, causing a fundamental problem when replacing a standard econometric toolbox with an ML toolbox. The success of AI in big data analytics results from their ability to uncover complex structures that may be unknown and completely specified in advance. ML models can be used to fit complex and very flexible functional forms to the data without overfitting; additionally, it can find available templates that work well out of sample. ML toolkits also scale well when applied to big data, which is not the case for some standard econometric techniques.

ML tools are new empirical tools to economics, and XAI attempts to address how the resulting prediction made with the AI tools correlates with known traditional (standard) techniques. Murdoch et al. (2019) refer to three "overarching desiderata for evaluation: predictive accuracy, descriptive accuracy, and relevancy" for the effectiveness of AI interpretability.

Model interpretability is domain-specific and an active area of research in econometrics. As a result, the general guidance provided in the economic literature recommends carefully deciding how to encode and transform the underlying variables when using ML techniques (Mullainathan and Spiess, 2017). The following section documents some of the current applications and economic research using XAI.

11.1.2 AI accountability models in economic research

Currently, the application of AI techniques in both economics and trade forecasting include the use of data collected from several different sources, such as a company's activity, news articles, and Twitter feeds, online reviews, and other sources, which may be very challenging for a human to monitor and in most cases impossible for a human to evaluate in real-time. Increasingly, the economics field has seen expanded use of AI and deep learning, taking advantage of the availability and accessibility of the tremendous computing power of graphical processing units (GPUs) in-house and in the cloud, resulting in AI hardware replacing humans in high-frequency trading (HFT).

A "black-box" model's decisions are sometimes too complex for a human to understand or may produce models that are challenging to troubleshoot. All the while, advanced AI methods, such as facial recognition, AI loan bots, AI interview evaluators, online recommendation systems, and more, are transforming everything about the economic world. For economic policy institutions to keep up, researchers need to provide transparency in the use of AI-driven models.

Additionally, in more regulated fields, such as health and criminal justice, explainable artificial intelligence (XAI) models have become crucial. This urgency is further driven by the need to reduce the risk of unintended consequences when employing these advanced solutions. Since AI-based algorithms already outperform humans, simply looking at the parameters of the model is beyond our understanding. Fortunately, AI researchers have found other ways to examine and understand an AI's output. This is where Explainable AI comes in; first is the interpretability (the ability to interpret an AI model), and second, the explainability (the ability to explain a model and its outcome in a human-centric way).

XAI methods can be split into two general categories: model-based and post-hoc (Murdoch et al., 2019). Model-based explainability refers to designing simple AI models, whose inner workings and decision logic can be easily represented and interpreted. Post-hoc XAI methods approximate "black-box" behavior by producing useful estimations of the model's inner workings and decision logic after receiving a trained and tested AI model as input. It then creates comprehendible representations in the form of the feature importance scores, rule sets, heatmaps, or natural language (Murdoch et al., 2019). There are numerous examples of XAI being employed to advance economics research, which will be covered in the next section.

11.1.3 Adopters of economic forecasting using XAI

Economists are employing machine learning applications not only to obtain better predictions, but also for policy decision-making. Batarseh et al. (2021a) suggest strategies to utilize when performing and developing AI assurance by providing a theoretical roadmap for a wide range of domains, including economics and finance. Recent economic-related AI modeling

attempts to incorporate performance indicators that can be analyzed and scored to both reflect a technology's human-centering and as a form of quality inspection. The leading economic forecasting research work that exemplifies AI assurance can be placed into three broad categories: forecasting, institutions, and finance.

Forecasting applications in microeconomics is still in its early stages. Yang et al. (2020b) present a class of interpretable neural network models that can produce highly accurate and interpretable predictions. The model encodes a type of interpretable function named persistent change filters, which allows the neural network to be written as a simple function of a standard number of interpretable features, indicating the outcomes of interpretable functions they have been encoded with. They then used the model to predict an individual's monthly employment status using high-dimensional administrative data, reporting close to 95% in-sample accuracy. The interpretable model performance compared favorably to some of the best conventional machine learning methods, whose prediction mechanisms are easier to understand. Severino and Peng (2021) evaluated fraud prediction in property insurance claims using various machine learning models based on data from a Brazilian insurance company. They estimated the impact of features in prominent cases of false positive and false negative model predictions using the explainable artificial intelligence methods SHAP and LIME. This explainable step would help risk analysts and professionals when working with AI-driven models for insurance claims.

Cook et al. (2021) use regression and PDP (one of XAI's model-agnostic tools) to estimate and examine the feature effects on the "slope and bootstrap variance" contributions of attributes of house pricing using the Iowa tax assessor dataset. The paper compares explanatory characteristics of OLS to that provided by PDPs when evaluating the feature effects of the direction and magnitude of changes in the predicted outcome as a result of changes in the feature values. The paper also compares LSE with Shapley and ALE; for each case, the interpretation is detached from and takes place after the model is fitted. Possible extensions to this research were provided to the motivated user, including an extensive comparison of OLS tables with Shapley values, examining links between instrumental variables

using causality-directed acyclic graphs (DAGs), and identifying feature interactions using available model-agnostic tools.

Historically, applications of ML have focused more on financial topics. Gramespacher and Posth (2021) noted the difficulties of using ML models to improve prediction accuracy, while following regulatory obligations, such as identifying the hidden cost associated with credit defaults in loan portfolios during the review and approval stages, while following regulatory obligations. The paper documents how ML methods can be used to address specific needs of credit assessment and how it makes sense to optimize for an economic target function rather than simply accuracy. Misheva et al. (2021) use LIME and Shapley values to explain ML credit scoring models developed using data from a US-based lending platform. LIME was used to explain local instances, whereas SHAP explained both local and global instances. Finally, the paper discussed some of the primary challenges associated with using these explainable tools on financial datasets, such as longer execution times, and suggested ways to overcome these challenges.

It can be difficult to interpret important features of AI models, though it is critical to have proper explainability when AI models are involved in influential aspects of life, such as banking. Hurlin et al. (2021) provide a framework for guidance on how lenders, controlled by their regulators, can monitor algorithmic fairness and improve model outcomes for the benefit of protected groups. Brusseau (2021) investigated if AI findings conform to that of humans, or vice versa, by modeling humanist investing in AI-intensive companies that are intellectually robust, manageable for analysts, useful for portfolio managers, and credible for investors. Carrillo et al. (2021) discuss the challenges of applying ML methods to real-world problems and how they affect relevant and novel explanations methods. Additionally, the article presents strategies to help mitigate these challenges when implementing explanation methods in the appropriate domain. The suggested mitigation strategies depended on the model-agnostic method used. Some of the most commonly used perturbation-based methods considered included PDP, ICE, ALE, SHAP, and LIME.

In the field of finance, there has been progress on using XAI. Ohana et al.'s (2021) article used a gradient boosting machine (GBM) tree-based ap-

proach to predict large S&P 500 price drops from a set of 150 technical, fundamental and macroeconomic features and reported an improved accuracy over other ML methods. They harnessed SHAP to identify the most important features to predict stock market crises, and subsequently selected a subset of those features for a final improved model. Their analysis uncovered the predictive role of the tech equity sector before and after the March 2020 financial meltdown. Giudici and Raffinetti (2020) present a global explainable AI model based on Lorenz decompositions and expands on previous contributions based on variance decompositions. They provided unifying variable importance criteria, which combined predictive accuracy with explainability, using normalized and easy to interpret metrics. The proposed decomposition was applied to predict bitcoin prices.

Economics XAI has even made progress in ways that will help economic institutions understand best practices. Navarro et al. (2021) provide a "user-centric desiderata" (real-world use cases of XAI to establish and address bias by ML-based decision-making models) discussing standard explainability requirements needed by statistical production systems at the European Central Bank. Preuss (2021) provides a review of the literature that details developments and the status of AI governance and the relevance of frameworks that ask for practicality, leading to suggestions of changes to regulations that are impossible to implement given current technology. Roa et al. (2021) discusses the impacts of alternative data, data collected from an application-based marketplace, contrasting traditional bureau data based on credit scoring models. Their results showed an improvement in predicting borrowers' behavior in groups traditionally underserved by banks and financial institutions, thereby validating the relevance of these datasets for predicting economic behavior in low-income and young individuals, who are most likely to engage with alternative lenders. Additionally, the paper discusses the use of the TreeSHAP method for stochastic gradient boosting interpretation. Their results revealed interesting non-linear trends in the variables originating from the app-based datasets, which may not usually be available to traditional banks. In their view, the results of their study present an opportunity for tech companies to disrupt traditional banking by correctly identifying alternative data sources and appropriately handling

the information they provide. Nesvijevskaia et al.'s (2021) article focus on fraud management issues and the intricacies of developing fraud detection models (FDM) to address the trade-off between accuracy and interpretability of detection. The article also provides a review of the different machine learning-based approaches to process fraud-related data. Finally, the paper examines ways to offer pragmatic and short-term responses to banks and policymakers without forcing economically and ethically constrained stakeholders into a technological race.

11.2 Commonplace explainability methods

AI models are in essence black boxes; it is difficult to understand how their immense complex connections come together to produce outputs. The main reason behind this difficulty are the many nonlinear relationships that exist across a large number of parameters or features. Currently, it is still hard to discern in concrete terms a global explanation of why an AI model is producing a certain outcome. Techniques and software are available in the research community to help address these concerns raised by AI accountability and include freeware visual tools to explain the datasets and models, data labeling tools, and commercial learning-based data annotation tools. These methods provide decision-makers a way to justify a model's behavior, which is critical for policymaking accountability. This section will discuss and implement some of the methods available for assuring ML models.

The two broad categories of explanations for predictive machine learning models are model-specific and model-agnostic. Freitas (2014) presented examples of model-specific explanations, including linear regression and logistic models (the simplest "native explainable" models), decision trees, classifiers such as SVM (support vector machines), and Bayesian probabilistic classifiers.

Model-agnostic approaches explain models in a post-hoc fashion by accepting the model as a "black-box," and then attempt to approximate their behavior (Wachter et al., 2017). An example of this approach is local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), which attempts to provide local explanations in the form of linear approxima-

tions of the model in small regions of the space. This approach is practical when explaining, for example, why a particular individual has been denied a mortgage application. Other post-hoc model-agnostic methods provide explanations in ranking features, even when the underlying model is not linear, include InterpretML (Nori et al., 2019), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and Partial dependence plots (PDP) (Friedman, 2001). Each takes a distinct approach to determining the important contributors to an ML model. The next sections will explain the methodology behind LIME, SHAP, and PDP and its implementation on the microeconomic cross-sectional dataset, called "The California Housing Dataset." This is a fairly common machine learning prediction dataset for California median housing (Pace and Barry, 1997).[1] Like the California housing dataset, the Boston housing dataset was also once included in scikit-learn's repository of pre-loaded datasets. However, it has been removed from future versions of the package, as it presents serious ethical problems for training data. The original goal of this dataset was to study air quality's effect on median housing prices. To do this, Harrison and Rubinfeld (1978) included features such as number of rooms, units built, etc., as well as one glaring feature that presents numerous ethical problems: the feature "B" referencing the "Black portion of the population" inappropriately identifies a potential causal relationship between race and Boston housing price elasticity. This model, in predicting housing prices, has encoded racism as a feature in predicting said prices, with little to no evidence that this is a factor in housing prices. Ethically sourced, nonbiased data is critical for training any AI model, so as a result, the California housing dataset will remain in the scikit-learn repository. In the code chunk below, the California housing data is loaded and previewed.[2]

```
# loading dataset
from sklearn.datasets import fetch_california_housing

df = fetch_california_housing(as_frame=True)
df.head()
```

[1] Data can be found here: https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html.

[2] Code explained in this chapter is on GitHub: https://github.com/AndersonMonken/AI-Assurance-Econ-Policymaking.

Table 11.1 Head of California housing dataset.

longitude	latitude	housing_ median_age	total_ rooms	total_ bedrooms	population	housholds	median_ income	median_ house_value	ocean_ proximity
−122.23	37.88	41.0	880	129.0	322.0	126.0	8.3252	452600	NEAR BAY
−122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500	NEAR BAY
−122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100	NEAR BAY
−122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300	NEAR BAY
−122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200	NEAR BAY

The output of the head function is shown in Table 11.1.

Subsequent sections will continue this example to conduct a random forest regression and use explainability methods on the model results.

11.2.1 Local interpretable model-agnostic explanations (LIME) explainer

LIME is a model-agnostic technique to explain the local environment around a single prediction (Ribeiro et al., 2016). LIME functions by explaining locally at an individual instance level, determining the differences in features, rather than assessing explanations at the aggregate or high model level. It attempts to fit a linear relationship for each individual production to explain why the instance produced a specific outcome. LIME provides the model with agnostic explanations making it easier to explain innumerable classifiers (such as random forests, support vector machines, and neural networks). LIME can explain the predictions of any classifier or regressor in a reliable way by approximating it locally with an interpretable model.

11.2.1.1 LIME methodology

Let f denote the original prediction model, and g denotes the explanation model, with $g \in G$, where G represents the class of potentially interpretable models (for example, linear models and decision trees). Suppose we target-specific local methods designed to explain a prediction $f(x)$ based on a single input x.

Then the explained model for x often uses simplified inputs x' that correlate to the original inputs through a mapping function $x = h_x(x')$. Local methods try to ensure $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$ with g acting as

input over absence/presence of the interpretable components (hence the domain of g is $\{0,1\}^d$.

As not every $g \in G$ may be easy to be interpretable, we denote $\Omega(g)$ to represent a measure of complexity (instead of interpretability) of the explanation $g \in G$ (e.g., for decision trees $\Omega(g)$ could represent the depth of the tree. For linear models, $\Omega(g)$ could denote the be the number of non-zero weights).

Next, denote by $f : \mathbb{R}^d \rightarrow \mathbb{R}$ the model being explained, by $\pi_x(z)$ a measure of the proximity between an instance z and x (a locality around x), and by $\mathscr{L}(f; g; \pi_x)$ a measure of g's approximation of f in local area π_x.

The local surrogate models that ensure both interpretability and local fidelity constraints can be expressed as minimization of $\mathscr{L}(f; g; \pi_x)$, whereas having $\Omega(g)$ is very low and ensures that the model is interpretable.

The interpretability constraint can be expressed mathematically as follows:

$$\xi(x) = \underset{g \in G}{argmin}(\mathscr{L}(f; g; \pi_x) + \Omega(g))$$

This equation can be used for different explanation families of G, fidelity functions $\mathscr{L}$, and complexity measures Ω. For example, in the case of a LIME explainer, one could focus on sparse linear models as explanations and perform the search around a particular prediction using perturbations.

11.2.1.2 LIME implementation

The package for LIME[3] has three main modules to use with different types of datasets:

1. ***lime_tabular***: used for structured dataset predictor explanations
2. ***lime_text***: used for textual dataset word explanations
3. ***lime_image***: used for image dataset pixel explanations

This section will cover lime_tabular. In a subsequent section, lime_text will be used to explain a sentiment result from a transformer network. The code chunk below shows the data being split into training and testing data at a

[3] LIME package: https://marcotcr.github.io/lime.

90/10 ratio as well as the random forest execution and the LIME explanation. The training performance is significantly better (97%) than the test performance (68%).

```
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from lime import lime_tabular

# preprocessing the data
df = df.dropna()
X = df.drop(columns=['ocean_proximity', 'median_house_value']).values
Y = df[['median_house_value']].values

# data must be a matrix for lime
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, train_size=0.90, test_size=0.1,
                                                    random_state=123, shuffle=True)
>> ((18389, 8), (2044, 8), (18389, 1), (2044, 1))

rf = RandomForestRegressor()
rf.fit(X_train, Y_train)
print("Test R^2 Score  : ", rf.score(X_test, Y_test))
print("Train R^2 Score : ", rf.score(X_train, Y_train))
>> Test R^2 Score  :  0.8071294626871858
>> Train R^2 Score :  0.9646086612627713

# lime method
explainer = lime_tabular.LimeTabularExplainer(X_train, mode="regression", feature_names= df.columns)

idx = 42
explanation = explainer.explain_instance(X_test[idx], rf.predict, num_features=len(df.columns)-1)
explanation.show_in_notebook()
```

Random forest is often one of the strongest machine learning models for structured data prediction. The LIME explanation is performed at the single instance level. The predicted value 222,290.00 is broken down into the variable's negative or positive effect, given specific rules; the values for that instance are given in Fig. 11.1. The values of each predictor are shown in Table 11.2.

11.2.2 SHapley Additive exPlanations (SHAP)

SHAP is a model-agnostic method that uses situational importance to measure a variable's importance (Lundberg and Lee, 2017). Unlike LIME, which approximates interpretable linear models in the area of a given prediction, SHAP draws from economic game theory, where each player receives a reward based on its contribution to the final outcome (Shapley, 1953). Shapley values describe the predictors contributions (i.e., additive contribution)

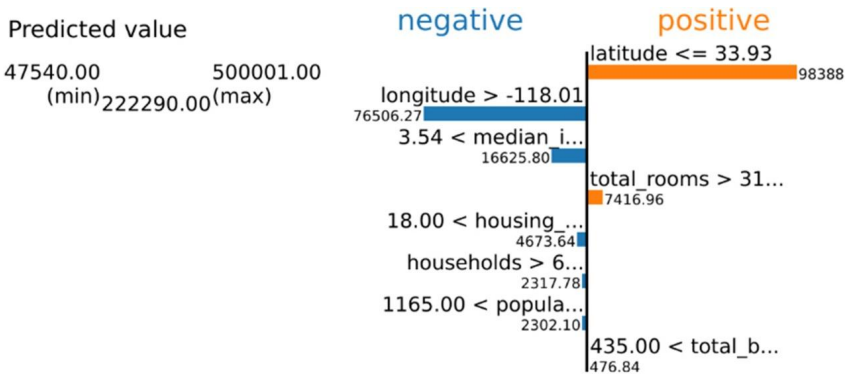


FIGURE 11.1 Python script that splits the California housing dataset, run random forest regression, and produce LIME explanation.

Table 11.2 Feature values for index 42 of the California housing dataset and positive (yellow (gray in print version)) or negative (blue (dark gray in print version)) effect on prediction.

Feature	Value
Latitude	32.78
Longitude	−117.07
median_income	4.71
housing_median_age	26.00
total_rooms	3725.00
Population	1516.00
Households	627.00
total_bedrooms	623.00

to an outcome and are used to explain why models make a particular prediction.

11.2.2.1 SHAP methodology

SHAP uses samples that provide estimates for feature importance in linear models in the presence of multicollinearity by computing how much variation the predictor promotes for a locally given point (how much a predictor contributes to its deviation from a given point). This is achieved by comparing a predictor's contribution to its expected contribution, as expressed

below, with β_j's depending on a subset of the input variables for a nonlinear case.

$$\phi_j(x) = \beta_j \cdot (x_j - E[x_j])$$

The interactions are measured by computing the contribution of variable j when fixed together with different subsets of variables.

To explain the working of SHAP, as with LIME, let with LIME, let f denote the original prediction model, N the number of features, and S the subset of features without our feature of interest i. Shapley values can be written as follows:

$$\phi_j(N, f) = \frac{1}{N} \sum_{Q \subseteq N \setminus \{j\}}^{\infty} (|S|!\,(|N| - |S| - 1))!\,(f\,(S\ U\ \{i\}) - f(S)\,,$$

where:

A. $1/N$ denotes each individual feature *i's* contribution to the overall prediction when it is included in the model

B. $\sum_{Q \subseteq N \setminus \{j\}}^{\infty}(|S|!\,(|N| - |S| - 1))$ denotes the permutation that will appear before and after i

C. $(f\,(S\ U\ \{i\}) - f(S)$ denotes the average using all features

The resulting value becomes the feature of interest's averaged marginal contribution towards the final prediction.

At a high level, the Shapley values conceptually aims to capture the importance and contribution by comparing the changes in the outcome for all possible orderings when the features of interest are introduced to a given model.

11.2.2.2 SHAP implementation

The SHAP implementation is conducted using the SHAP package in python.[4]

The same random forest regression is used from the previous explainability implementation, as shown in Fig. 11.2.

[4] SHAP package: https://github.com/slundberg/shap.

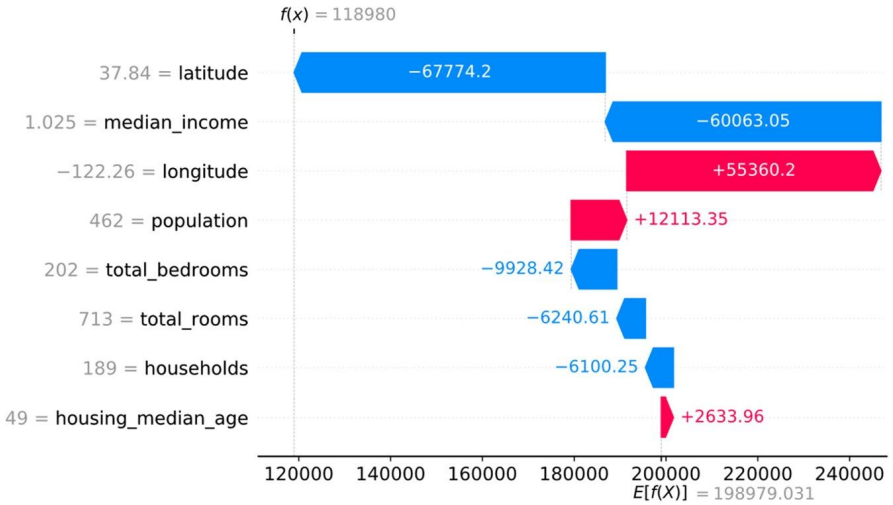


FIGURE 11.2 Python script that runs SHAP explainability method on random forest prediction.

```
# preprocessing the data
df = df.dropna()
X = df.drop(columns=['ocean_proximity', 'median_house_value'])
Y = df['median_house_value']

from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, train_size=0.90, test_size=0.1,
                                                    random_state=123, shuffle=True)
>> ((18389, 8), (2044, 8), (18389, ), (2044, ))

# SHAP to explain feature contribution
explainer = shap.Explainer(rf.predict, X)
shap_values = explainer(X)

# sampling 100 SHAPley values
X100 = shap.utils.sample(X, 100)

# feature importance plot
idx = 42
shap.plots.waterfall(shap_values[idx], max_display=14)
```

The SHAP explainer perturbs the model to find out the Shapley values for each variable. This process can be computationally intensive for large models or datasets. The waterfall diagrams include rich information about the same instance as LIME explained. The random forest $E[f(x)]$ or average prediction is 22.326, and each of the variables listed are given credit for

changing from the baseline. The final prediction is 24.98, and the strongest variables are "LSTAT", "RM", and "CRIM".

11.2.3 Partial dependence plots

Dependency visualizations often provide a way to gain intuitive insights into the relations that exist in a dataset. Partial dependence plots (PDP) provides a model-agnostic method of showing relationships between feature and predictor variable (Friedman, 2001).

PDP averages the results for the features that we're interested in (typically one or two) across a specified range of values, in effect marginalizing the other features to determine the dependence of the specific features of interest. PDP plots are then obtained by plotting values of the feature of interest by the averaged predictions. If one feature is considered, the PDP plot layout includes feature values on the x-axis and the prediction on the y-axis. While two features of the PDP plots look like a heatmap with the two features on the x and y axes and the color scale as the prediction. By observing trends from the PDP, one can understand the behavior of features in a prediction model. As implemented in Cook et al. (2021), PDP serves as an analog to coefficient from a traditional econometric model.

11.2.3.1 PDP methodology

To explain the working of PDP, as with LIME, let f denote the original prediction model, and partial dependence function f_{xs} is the model using our features of interest, x_s, and x_c are the remaining features, consisting of our full feature space. Finally, we let $P_c(x_c)$ be the probability density of x_c, so we have

$$x = x_s \text{ U } x_c$$
$$f_{xs}(x_s) = \int f(x_s, x_c).P_c(x_c).dx_c$$

As it is not feasibly to integrate over all possible values of x_c, the integral can be estimated by taking the average over a given dataset, simplifying the

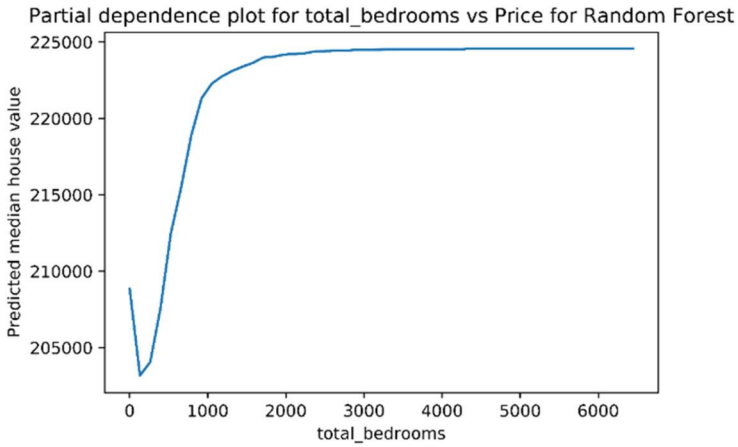


FIGURE 11.3 Python script that runs the random forest regression and produce partial dependence plot.

given equation as

$$f_{\mathrm{xc}}(x_{\mathrm{s}}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{N} f\,(x_{\mathrm{s}}, x_{\mathrm{c}} i)$$

Conceptually, we hold the features of interest x_s constant and compute predictions over all other combinations of the complement set x_{c} by averaging out predictions across the given dataset to obtain the partial dependence value for that instance.

A plot (PDP plot) can then be generated after computing averages for each instance over a specified range of values in x_{s}; a plot (PDP plot) can then be generated.

11.2.3.2 PDP implementation

In Fig. 11.3, a linear vector is created that spans the entire range of the predictor of interest "total_bedrooms." While holding other predictors constant, the value of "total_bedrooms" is stepped from its minimum to its maximum, while the model is re-predicted using the modified input predictors. The average prediction is taken for each step of the "total_bedrooms" predictor and plotted to show the prediction's partial dependence on the

"total_bedrooms" value. It is also possible to bootstrap at each step to produce a bootstrap confidence interval across the "total_bedrooms" values.

```
# pdp method
rf_model = RandomForestRegressor(n_estimators=100).fit(X_train, Y_train)

total_bedroom_values = np.linspace(np.min(X['total_bedrooms']), np.max(X['total_bedrooms']))

pdp_values = []
for n in total_bedroom_values:
    X_pdp = X.copy()
    X_pdp['total_bedrooms'] = n
    pdp_values.append(np.mean(rf_model.predict(X_pdp)))

plt.plot(total_bedroom_values, pdp_values)
plt.ylabel('Predicted median house value')
plt.xlabel('total_bedrooms')
plt.title('Partial dependence plot for total_bedrooms vs Price for Random Forest')
plt.show()
```

11.3 Mitigating bias in AI models for economic prediction

The first opportunity to mitigate bias in the AI pipeline is in the pre-processing stage, where the data cleaning, wrangling, and feature engineering steps occur before any algorithm is applied. Thus pre-processing is independent of the algorithm/model itself.

One of the mitigation techniques to use at this stage is reweighting. This involves assigning weights row-wise to existing biased training data based on the frequency of said bias in the dataset. The main advantage of this approach is that the underlying data is left unchanged, while bias is mitigated (Kamiran and Calders, 2017). However, it is important to note that this technique is not involved in the model training process, and therefore some accuracy-discrimination tradeoff is expected.

One of the most well-known and widely used datasets in the machine learning world is the adult dataset gathered from the 1994 US census and is now publicly available on the UCI machine learning repository.[5] Table 11.3 shows same dataset restricted to the gender and income categories. The probabilities represent the portion of a subgroup under a particular income category. The primary goal in this dataset is predicting whether an individual has an income higher than $50k. Male is the privileged group

[5] Find the data here: https://archive.ics.uci.edu/ml/datasets/adult.

Table 11.3 Adult Dataset (Restricted to Gender and Income Counts)[a].

Gender	Income	Count	Probability
Female	<=50k	9592	89%
Female	>50k	1179	11%
Male	<=50k	15128	69%
Male	>50k	6662	31%

[a] Similar study can be found here: https://towardsdatascience.com/reweighing-the-adult-dataset-to-make-it-discrimination-free-44668c9379e8.

(31% probability of having a positive outcome, income greater than $50k), whereas female has an 11% probability of this outcome. Also, it is worth noting that this dataset is limited in its scope of gender identity.

Before manipulating the data directly, it is imperative to think about how to quantify this idea of unbiasedness (or fairness). The disparate impact metric is one of these measures (Cesaro and Gagliardi Cozman, 2020). It is found by taking the ratio of the sensitive group with a positive outcome, which in this case is females with an income greater than $50k over the non-sensitive (privileged) group with a positive outcome (males with income greater than $50k). A score of 1 indicates the dataset is "completely unbiased," meaning that both are equally likely to achieve this outcome.

$$DI = \frac{P_{positive\ \&\ unprivileged}}{P_{positive\ \&\ privileged}}$$

Equivalently,

$$DI = \frac{P(\hat{Y} = 1 | A = female)}{P(\hat{Y} = 1 | A = male)}$$

Weights should be assigned according to the frequency counts from Table 11.3. For example, for the female (unprivileged) group with an income of under $50k (negative outcome), the weight is calculated as follows:

$$W_{negative\ \&\ unprivileged} = \frac{\#\ of\ unprivileged * \#\ of\ negative}{all * \#\ negative\ \&\ unprivileged}$$

$$W_{negative\ \&\ female} = \frac{10{,}771 * 24{,}720}{1{,}256{,}257 * 9592} = 0.8525$$

The authors use reweighting on the entire adult dataset (as well as others) and test it on a variety of models. They apply reweighting and calculate the DI score and equality of opportunity metrics to their models using the AIF-360 library by IBM.[6] The weights, calculated as indicated above, serve as inputs to their logistic regression model. Introducing reweighting at the beginning of the machine learning pipeline results in a much higher DI score, slightly above the 0.8 DI score specified by US employment law. They also utilize SHAP to explain individual predictions. Their results show that the introduction of reweighting favors the unprivileged group, as expected, significantly reducing bias.

In-processing bias mitigation describes interventions made during the learning process of the model. Adversarial debiasing is one of these methods, and it involves building two models. The first model aims to predict the target variable, based on whatever feature engineering steps that have already been taken. The second model tries to predict, based upon the predictions of the first model, the attribute contributing to that bias (gender, race, etc.). In an ideal world, the second "adversarial" model should not be able to predict the sensitive attribute, indicating virtually no bias. This model therefore guides modifications to features and parameter weights in the predictor model to weaken the predictive power of the adversarial model until it cannot predict the sensitive attribute. While originally proposed by Goodfellow et al. (2014) in the context of generating images, this adversarial learning framework is generalizable to any model in which one has access to model parameters.

Zhang et al. (2018) proposed a framework for a classifier trained on the adult dataset to predict income. They considered the case where gender is both implicitly and explicitly included in the data. They implemented their first model which aimed to predict Y based on X, while modifying weights W to minimize model loss L_p. They then defined their adversarial model to predict sensitive attribute Z from $\hat{Y}$. More formally, the modification rule

[6] Package is available for both Python and R https://aif360.mybluemix.net/.

for the weights W of the model is

$$\nabla \mathrm{W} L_p + proj_{\nabla \mathrm{W} L_\alpha} \nabla \mathrm{W} L_p - \alpha \nabla \mathrm{W} L_\alpha$$

where α can be changed depending on the desired balance between model accuracy and unbiasedness.

Using this framework, the authors trained two logistic regression classifiers, one with mitigation and one without, and found that this adversarial learning framework for predicting income resulted in near equality of odds. More formally, $\hat{Y}$ and Z are independent given Y, meaning the predictions are uncorrelated with the sensitive attribute, which was their goal. They found that this adversarial debiasing framework results in a small 1.5% decrease in accuracy, and that the false positive and false negative rates are approximately equal across subgroups. Assuring that models are not biased does not necessarily mean degraded performance.

Post-processing occurs after the model has been trained and used to make predictions, and, like pre-processing, doesn't require access to the inner-workings of the model, making this stage of bias mitigation suitable for any AI method. Calibrated equalized odds is a post-processing algorithm proposed by Pleiss et al. (2017), in which a cost function is introduced to penalize the model for disparities in false negative rates, false positive rates, or a combination of cost functions across sensitive groups, such as gender groups, racial groups, or age brackets. Equalized odds is achieved when the sensitive and unsensitive groups (in our case, gender) have equal error rates, according to some cost function of choice. This is to ensure that an error type doesn't disproportionately impact one group over another. The choice of cost function is dependent on the specific problem at hand. Using the same 1994 adult dataset, the authors harnessed the false negative condition to predict income with equalized costs across genders. Modifying the outputs from their logistic regression model, which outputs probabilities, the authors adjust the probability thresholds for sensitive groups such that balance is achieved for calculating false negative rates across both groups. Introduction of this cost function to mitigate bias between genders decreases the overall accuracy rate by 10%.

The aforementioned examples are a non-exhaustive list of mitigation techniques and frameworks. IBM's AI 360 Fairness 360 is an open-source toolkit with bias mitigation tools and metrics for any processing step, including those mentioned above. Feldman and Peake (2021) studied the latter two mitigation frameworks and concluded that their *end-to-end bias mitigation approach*, in which they employ bias mitigation techniques at every stage of the project lifecycle, proves to be more effective than any singular mitigation technique at maintaining fairness across multiple metrics.

11.3.1 NLP use in central banking

Data from social media, earnings calls, policymaking meetings, and more, have become publicly available over time, opening many new avenues of opportunity for data scientists to study our economy. This increase in data availability, specifically public access to economic information, such as data, models, and forecasts relevant to the central bank's decisions, reflects positively on the Federal Reserve System and the European Central Bank. According to a study conducted by European Central Bank economists, the abandonment of hidden discussions concerning monetary policymaking has ushered in a new era of transparency, which both reduces inflationary biases and gives the central bank more flexibility to respond to economic shocks (Geraats, 2002).

The Federal Open Market Committee, consisting of the seven Board of Governors members and five of the 12 Reserve Bank presidents, is the primary monetary policymaking body of the Federal Reserve system. Eight meetings are scheduled annually, where the committee addresses the most relevant market information disseminated in statements and minutes. NLP can be specifically employed to study the impacts of policy changes on financial outcomes.

Though ample data is available on the financial market effect of FOMC meetings and utilizing natural language processing to predict stock prices, there are few rigorous studies on the effect of FOMC minutes on asset prices. Carlo Rosa[7] studied the effect of FOMC minutes releases on treasury rates, stock prices, and U.S. dollar exchange rates over a six-year time

[7] https://www.newyorkfed.org/medialibrary/media/research/epr/2013/0913rosa.pdf.

period. She found that the release of these meeting minutes has a tangible impact on asset prices. When minutes are released, two-year treasury yields suddenly increase roughly three times larger on event days than normal (non-FOMC release) days. Treasuries at shorter maturities are the most affected asset class, followed by U.S. dollar exchange rates. However, the asset price response, or any changes in asset price immediately after the release of FOMC minutes, has declined since 2008. This decline reflects more transparent and consistent discussion of the committee. FOMC discussions have been clear in their objectives, such that policy discussions over time should not have shocking results, as shocking results contribute to increased volatility. The decrease in asset price volatility after FOMC minutes releases suggests the effect of greater transparency on behalf of the FOMC. Economic policymakers and data scientists alike should be aware of these asset price effects, as central bank communications have clear and tangible effects on financial outcomes. FOMC meeting data must be handled sensitively, because the information contained in their release can have profound market effects.

NLP techniques allow us to predict producer/consumer expectations, as well as identify rapidly changing trends during black swan events. Data from quarterly earnings calls, where a firm discusses their financial progress and results for a given period, have become increasingly more publicly available, giving data scientists a new avenue to study producer expectations. While data from news articles and social media provided insight into consumer expectations for the economy, Hassan et al. (2020) monitored firms' earnings calls transcripts during the beginning of the COVID-19 pandemic to identify adverse or positive effects on firms. A global phenomenon requires global coverage of data, and the authors collected data on both domestic and international firms. Checking these outcomes at the sub-group level is critical for confirming that Simpson's paradox[8] does not affect the overall firm results (Simpson, 1951).

Hassan et al. (2019, 2020) first identified which portions of the transcripts contained COVID-19 and other disease-related discussions, which

[8] Simpson's paradox is when the overall trend in the data is not representative of individual sub-group trends.

they did by comparing training libraries with disease-related text to non-disease-related text. They began by taking a list of pandemic diseases updated on the World Health Organization website and focused on outbreaks within the sample period beginning in 2002 and ending in March 2020, and then restricted their training list by only including diseases that attracted an international audience and were concerning to investors. They then included synonyms to the remaining list of outbreaks to ensure that all disease-specific related information is picked up from transcripts. Fine-tuning speech-to-text data, as in the case of earnings calls transcripts, requires an additional layer of assurance that captures colloquialisms as well as synonymous terms used in speech over text. For example, "novel coronavirus," "COVID-19," "coronavirus," and "covid" can be used interchangeably (depending on time-varying context), but whether we are analyzing formal text form the NIH or a phone call between companies, one name for the disease may be more prevalent than the others, so it is important to account for all of them.

In this example, a lot of training data related to "coronavirus" is time dependent. Coronaviruses have existed for centuries and have been studied for decades. Though this case may be particularly obvious in identifying that the coronavirus discussed in earnings calls is the COVID-19 that has impacted every corner of the globe in recent years, this principle of assuring data in the appropriate context pervades every domain, just as assuring facial recognition technology recognizes a multitude of skin tones in a variety of lighting environments. The research group conducting this study also performed a human audit on a sample of transcripts to ensure that they were using the correct word combinations for the associated disease outbreak and verified that these combinations had no alternate meaning other than the disease in question.

By analyzing the presence of these disease-related "bigrams" (two-word combinations) in light of firms' business operations, the authors were able to identify the top three issues firms were facing: supply chain disruption, decrease in consumer demand, and employee welfare. Some of this analysis also highlighted business opportunities for certain firms specializing in testing equipment, antiviral medication, etc. As the first quarter of 2020

progressed, an increasing number of firms were expressing concerns about employee welfare, especially in the context of work from home options, and supply chain discussions nearly tripled.

Hassan et al. (2020) also considered firm-level responses in the context of past diseases, SARS and H1N1 in particular. Does prior exposure to diseases help firms deal with the gravity of the COVID-19 pandemic? To analyze how well firms have dealt with COVID-19 in the context of past diseases, the authors conducted coronavirus sentiment analysis in the context of prior epidemic exposure. They found that firms that had more extensive discussions around SARS and H1N1 had significantly fewer negative sentiment scores related to COVID-19, suggesting a more positive outlook on handling this pandemic compared to those who haven't dealt with diseases in the past.

But how are these sentiment scores determined? Though the paper includes an OLS regression to predict negative COVID-19 sentiment and counts the use of negative-tone words used in conjunction of these discussions, it does not consider the subject of these negative tone words. Sentiment analysis is seemingly a black box of words to scores, so transparency on how those scores are determined and calculated is imperative in NLP research. Whether sentiment scores should be calculated through a transformer model, such as "Bidirectional Encoder Representations from Transformers" (BERT) or a sentiment dictionary, such as VADER, is up to the data scientist, but it is important to assess the ramifications of both these different methods.

Dictionary-based approaches are among the simpler of the two methods, involving creating two dictionaries, of terms (words or phrases) carrying positive and negative sentiment, respectively. If a text hits both dictionaries, it is classified as both, and classified as neutral if it hits neither. This approach would prove beneficial if one were to analyze smaller strings of text (like short responses in a chat, as opposed to long earnings calls transcripts) and would also not run into any of the issues that arise by using machine learning models in general for sentiment analysis.

11.3.2 NLP transformer networks

Whereas sentiment dictionaries rely on a simplistic and straightforward implementation, BERT and its varieties employ a transformer language model, a mechanism implemented using convolutional neural networks trained with attention models at each stage. BERT is also trained on an incredibly large corpus of text, which includes the entirety of Wikipedia and Book Corpus. The advantage of the BERT model is that it is bidirectional (the B of BERT), able to read an entire sequence of words at once (not just from left to right or right to left), allowing the model to learn the context of a word based on its surroundings. Spoken language is incredibly complex, teeming with variations in tone and use of metaphors and colloquialisms. Since much of NLP applied to economics involves analyzing transcripts of spoken dialogue, it is imperative that the models used to analyze this text are robust in understanding the relationships among words non-sequentially and over the span of a large corpus of text. These corpora, such as central bank chairman speeches, firms' earnings calls transcripts, or even press conferences, are far too extensive to rely on simple sentiment dictionaries that do not capture long term relationships/dependencies between words and ideas.

In the case of analyzing FOMC minutes and statements, the complexities of these texts span across time periods. Simply knowing whether or not we have reached a "positive" term (via sentiment dictionary) won't give any context. To notice policy trends, we must gather and analyze this data over time, and capture long-term relationships within the text. Though it may seem that BERT is the clear winner in the large-scale sentiment analysis category, it is costly. This trade-off between model complexity and computational resources is an essential consideration a data scientist must make when deploying any model.

Sentiment varies greatly depending on its domain or context. For example, "Significant upside profit potential" has a far more positive sentiment than "Heighted upside risk to inflation outlook." As a result, these general-purpose models are not nearly as effective in conducting specialized sentiment analysis as their "finely-tuned" counterparts. Using language models that are pre-trained for specific circumstances has become

more commonplace in natural language processing. As the applications of NLP become more specific, task-specific modifications to training models are required. As in the COVID-19 earnings calls case study, financial discussion utilizes specialized language. Assuring that the corpus of data used to analyze a COVID-19-firm, related text is fine-tuned for financial discussion is a means of assurance. FinBERT, a language model based on BERT, is a financial domain-specific transformer model that outperforms other robust machine learning methods in financial sentiment analysis (Araci, 2019). It is trained on years of earnings calls transcripts, analyst reports, and the similar texts.

However, the underpinnings of BERT, let alone FinBERT, can be quite complicated to understand for someone who is not a machine learning engineer. Its sheer power to analyze and classify text requires substantial transparency between itself and the user. Explainable Artificial Intelligence tools, such as LIME and SHAP, are incredibly useful here. LIME is primarily used to explain models that classify data by category (examples include facial recognition, sentiment analysis, etc.) via learning a local linear representation around the prediction.

11.3.3 LIME for text explanations

Consider a finance-related tweet uploaded to the IEEE dataport (Taborda et al., 2021). The pretrained FinBERT model can be used with an appropriate tokenizer, which breaks up texts into small chunks, or tokens, for analysis (Araci, 2019). All these tools are open source. After loading the model and data, one can perform sentiment analysis on a single tweet shown, as shown in Fig. 11.4.[9]

The standard output for sentiment analysis using BERT is SoftMax activations. This is often used as the last step in a neural network, where a vector of numbers is converted to a vector of probabilities, normalizing these outputs so that they are interpretable as a probability for each class.

[9] Dataset publicly available here: https://ieee-dataport.org/open-access/stock-market-tweets-data.

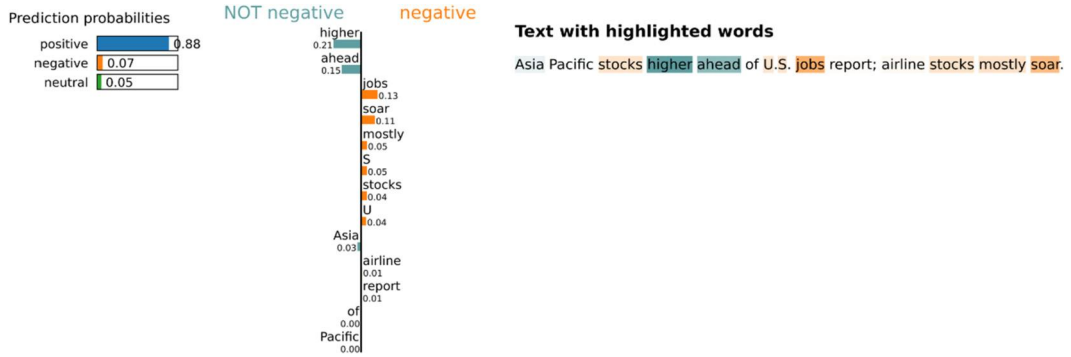


FIGURE 11.4 Python script that explains sentiment prediction - Feature Importance for Tweet.

This tweet is classified as positive, with a probability of 88%, though it would be insightful to know why. What specific attributes contribute to this result? LIME is one method to explain the probability outcome.

```
text = 'Asia Pacific stocks higher ahead of U.S. jobs report; airline stocks mostly soar.'

# Breaking the text into small chunks for analysis
tweet_batch = tokenizer(text, padding=True,truncation=True,max_length=160, return_tensors="tf")

# Putting the text through our model
tweet_outputs = model(**tweet_batch)

# Logit output
tensor_logits = tweet_outputs.logits

# Final probabilities
tweet_prediction = tf.nn.softmax(tensor_logits, axis = -1)
tweet_prediction
>> <tf.Tensor: shape=(1, 3), dtype=float32, numpy=array([[0.87840664, 0.06923903, 0.0523543 ]],
dtype=float32)>

# Returns the index of the highest probability
index = np.argmax(tweet_prediction.numpy())
index
>> 0

# Indexing into the sentiment classes to return the appropriate label
sentiment_classes[index]
>> 'positive'
```

The LIME explainer uses a function that takes in a string of text and the pretrained FinBert model and outputs a list with the effect from each word in the input text, as shown in Fig. 11.4. LIME is what is known as a post-hoc technique, as the explainability tool is introduced "after the fact" of model execution and not from the beginning of preprocessing and model training

(ante-hoc). After the model is trained and run on a string of text, LIME's inner workings allow us to see which attributes of the text contribute most to the overall positive result. It does this by iteratively perturbing the input, the tweet, with token masking, and then calculating sentiment scores at each iteration. For example, LIME would remove the word "higher" from the tweet and calculate the sentiment score, which would be relatively lower than the result from using our full tweet. Comparing the two scores shows the impact that single feature (word) had on overall sentiment.

```
from lime import lime_text
from sklearn.pipeline import make_pipeline
from lime.lime_text import LimeTextExplainer
import matplotlib.pyplot as plt

explainer = LimeTextExplainer(class_names=sentiment_classes)
exp = explainer.explain_instance(text, predictor, num_features = 20,
                                 num_samples=2000)
exp.show_in_notebook(text=text)
```

Seeing how sentiment scores change with the removal of particular features (words) gives the explainer an idea of which features are more important. The list of attributes consists of single words, and you may remember that models, such as FinBERT, understand both long term dependencies and relationships between words in a string of text to output their results. Though it would be favorable to see which word relationships contribute to this, it is important to remember LIME approximates this model linearly, so it may not be possible to uncover these black-box relationships entirely.

Once again, the nuances of language are complex, and these models only go so far in mimicking the human intelligence that NLP strives for. Many language models have yet to find a consistent and accurate way of detecting sarcasm and analyzing metaphors. Though these models are used in primarily formal settings in the context of economic policymaking, likely devoid of much sarcasm, it is critical to understand the nuances of the domain of study before deploying a language model. Policymaking accountability depends on it.

Though transparency is of utmost importance as models become more scalable and ubiquitous, assurance is needed in all parts of the project lifecycle. Data collection, though seemingly mundane, is pivotal in shaping

economic machine learning outcomes. It is important to think critically about the data samples being gathered and used to train a model. If the sample consists of a bunch of web-scraped finance-related posts from Reddit and Twitter, then one must assure the quality of the text and confirm it is free of bias. The training process is where the "learning" in machine learning takes place. Let's say these Reddit posts and Tweets frequently put "female" next to "teacher" and "male" next to "professor." This bias could permeate into a model. Word embeddings allow data scientists to turn words into a series of numerical values to then input into machine learning models. These word embeddings still contain the aforementioned patterns, and the model will learn those biased patterns and perpetuate the associations of female: teacher and male: professor. Online texts are filled with human stereotypes, expressed explicitly and implicitly; these stereotypes, or biases, present in data are amplified by machine learning models. Why?

It is no secret that the world is teeming with inequality: the gender salary gap, racial disparity in homeownership, etc. Data from a historically unequal society perpetuates those existing inequalities if not dealt with directly. Inequalities are present in the dialogue used online, whether consciously expressed or not, and therefore layers of assurance are required at every step of the project lifecycle.

11.3.4 LLMs and the AI central banker

Two generative models that have defined the LLM landscape are generative pre-training transformer (GPT) models and bidirectional encoder representations from transformers-based (BERT-based) models. There are many more LLMs that have accelerated the improvement of LLM performance, such as ERNIE, ELMo, GROVER, Big BIRD, Rosalita, and many others, as shown in Table 11.4. However, the GPT and BERT projects capture both the immense power of LLMs and the associated risks that come with them.

The GPT project by OpenAI highlights the importance of model parameter size in creating human-like text. There have been three iterations of the GPT project so far, each increasing the number of features within the model exponentially. One of the primary goals of this project is to create a robust unsupervised zero-shot model that requires little fine-tuning to execute a

Table 11.4 Current Most Impactful LLMs.

Timeline	Model	Institution	Assurance Considerations	Parameter Size	Citation
2018-2020	GPT-1, 2, 3	OpenAI	Pre-trained on BookCorpus; very human-like outputs; GPT-1 and GPT-2 are open-source; English language dominate	117M; 1.5B;175B	Radford et al. (2019); Brown et al. (2020)
2019-2021	ERNIE 1, 2, 3	Baidu	Trained on text and graph data; pre-trained on 4TB dataset; not publicly available; Chinese language dominate	114M;10B	Sun et al. (2019a); Sun et al. (2019b); Sun et al. (2021)
2020	T-NLG	Microsoft	Pre-trained on unknown dataset; not publicly available; cited to be integrated into Microsoft products	17B	Rosset (2020)
2018	BERT	Google	Pre-trained on BookCorpus; foundation for many other LLMs; open-source; used in many Google products	340M	Devlin et al. (2018)
2019	Megatron-LM	Nvidia	Pre-trained on OpenWebText; removes non-English content; pyTorch model parallelism	8.3B	Shoeybi et al. (2019)
2020	XLNet	Google	Auto-regressive model; unidirectional nature; not pre-trained; English language dominate	110M	Yang et al. (2020a)

variety of tasks. GPT models initially develop most of their task understanding from a thorough unsupervised pre-training process that is reliant on the contents of the pre-training dataset. These models are then trained again

in a supervised fine-tuning process to understand the nuances of a specific topic or task with a training dataset.

In 2018, the GPT-1 project created a proof-of-concept model that showed the utility of semi-supervised learning in LLMs. Radford and Narasimhan (2018) started with a 117M parameter unsupervised model that was trained on a large textual dataset, called BookCorpus. Then the model underwent a supervised fine-tuning process that included an auxiliary learning objective focused on language modeling that was optimized along with its primary objective. In other words, though the GPT model's primary objective was predicting word sequences, it also optimized NLP tasks besides prediction. Model development can be shown in the equations below, where t is the given unsupervised corpus of tokens, and k is the context window size for conditional probability using θ parameters. The second equation illustrates the objective maximization of x tokens for y labels. The third equation shows the influence of the auxiliary objective optimization $L_1(C)$ with a λ hyperparameter weight.

Auxiliary unsupervised language modeling objective:

$$L_1(T) = \sum_i log\, P(t_i \mid t_{i-k}, \ldots, t_{i-1}; \theta)$$

Primary supervised fine-tuning objective:

$$L_2(C) = \sum_i log\, P(y \mid x_1, \ldots, x_n)$$

Combined loss function with primary and auxiliary objectives:

$$L_3(C) = L_2(C) + \lambda L_1(C)$$

The unsupervised language modeling auxiliary objective technique was first implemented in an earlier study by Rei (2017), and OpenAI improved upon its effectiveness with its pre-training process when it concluded that the technique was most beneficial for large datasets. This model update resulted in two important outcomes as highlighted in the seminal GPT paper: improved generalization and faster convergence for LLMs. Better generalization for LLMs meant that GPT-1 was able to work with unseen data more

easily when the data was from the same distribution of data it was trained on. Faster convergence revealed that it took less examples from a distribution for a model to understand it, which made the model more accessible over a diverse set of tasks. Overall, the model's improved generalization capabilities translated to successful zero-shot setting performance compared to state-of-the-art models. LLM performance is typically assessed by looking at how well an LLM completes a set of tasks when trained on a challenge (benchmark) dataset. When comparing task performance on NLP tasks (such as question answering, commonsense reasoning, semantic similarity, and classification) between GPT and specifically trained models, it was reported that GPT performed the best on 9 out of 12 challenge datasets.

The next iteration of the project in 2019 was GPT-2, which focused more on model parameter size compared to GPT-1. Updates to the training methodology for GPT included adding task conditioning to increase zero-shot performance, increasing model parameters to facilitate more generalization, and switching to a new pre-training dataset, called WebText (Radford et al., 2019). The model was split into four sizes (117M, 345M, 762M, and 1.5B) to better identify the influence of model parameter size on NLP task performance. It was found that as the number of parameters in the model increased, the perplexity of tasks decreased regardless of the training dataset. This essentially meant that the model's prediction error measure for the primary objective decreased as the features available to the model increased. The outcome of this was that across NLP tasks, the model with the highest number of parameters (1.5B) performed the best out of all the parameter sizes of GPT-2. Unsurprisingly, the largest version of the model outperformed most existing unsupervised models, similar to the results of the GPT-1 assessment. Despite the continued success of the GPT project, it was noted that GPT-2 underfit the WebText dataset, suggesting that the size of these models was still too small to capture all the trends in the dataset.

In the realm of economics, zero-shot or few-shot capabilities in LLMs are important for applications due to a basic issue that influences all language processing tasks, the lack of well-labeled-content-specific big data. Properly fine tuning an LLM requires significant investment in hardware, software, and time to build a labeled training dataset. Open source tools exist

for labeling that make the process easier, such as Doccano, which is a scriptable annotation tool for machine learning.[10] Even with Doccano, however, the number of labels needed for a robust training process is still in the thousands or millions, even to execute tasks, such as generating short economic summaries. It would be much easier for an analyst to use a model that can perform tasks well enough without the need to fine-tune, but LLMs haven't reached this level of performance yet. Currently, an analyst or economist trying to utilize these models would need to dedicate immense resources to build a training dataset with proper labels prior to the fine-tuning process. Therefore the first two GPT models and the majority of LLMs are not accessible to most researchers studying economics since the development of datasets require great effort.

The most recent iteration of OpenAI's experiment with GPT language models is GPT-3, released in 2020, which is currently the largest neural network ever created. The most impressive aspect of the latest model is how simply dramatically increasing the parameters from 1.5 billion to 175 billion made output incredibly diverse and human-like (Brown et al., 2020). The large capacity of this model made writing essays and responses less distinguishable to a human's writing, raising red flags on the potential public use of this model. Important results from this model is its drastically improved quality performance on few-shot setting tasks compared to zero-shot settings and its improved fine-tuned in-context learning. Ultimately, the latest release of GPT-3 has been a great leap forward in LLM development that was almost entirely related to its immense size. However, GPT-3 is not an open-source project like its predecessors, making use of it for economic research even harder to explore.

The most obvious use for these powerful LLMs in macroeconomics is to predict the language of future monetary policy statements, an idea that is already being tested in the private sector on FOMC statements. A reasonable goal that an institution can achieve with LLMs is fine-tuning a model to draft and predict the language of future monetary policy statements across major economies, while also classifying these texts along a hawkish or dovish response gradient, as illustrated in Fig. 11.5.

[10] Doccano code is available here: https://github.com/doccano/doccano.

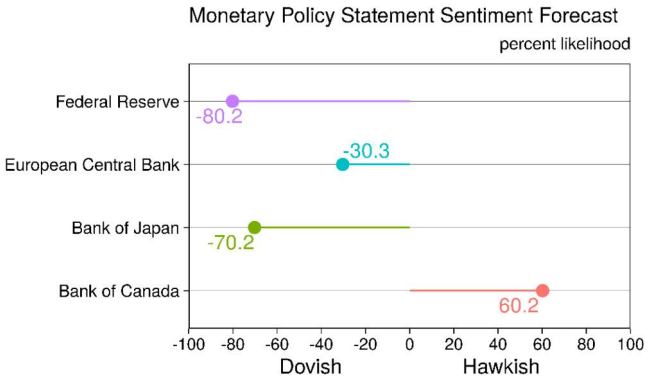


FIGURE 11.5 This figure is a visualization of the possible likelihood of sentiment of upcoming monetary policy statement releases across several central banks. This prediction can be created by analyzing sentiment on generated central bank statements that have been produced when providing a large language model (LLM) macroeconomic indicator conditions. This would come from a fine-tuned LLM that has been trained on various central bank monetary policy statements and has understood language logic related to historical events.

A project to predict monetary policy text would require a robust dataset of monetary policy statements, press conference responses, and various other examples of how a central bank communicates its policy intentions to the public. Most of this data is offered publicly on central bank websites and on other public online resources. The more predictable the syntax, terms, and connotations of the textual policy examples and the larger the dataset, the easier it would be for LLMs to adopt a reasonable response pattern when given new information. As recalled before, LLMs function by predicting the most likely reasonable response text when provided with a task, such as answering a question. Therefore, a well fine-tuned LLM model could provide policy recommendations when given specific economic conditions that are generated from previously applied policy actions within the training textual dataset.

A study by Middlebury Institute of International Studies showed evidence of GPT models and other advanced neural language models inheriting the ideology of focused datasets (McGuffie and Newhouse, 2020). For economics, this means if an LLM was trained on discriminating between "hawkish" and "dovish" responses, it could further adapt its textual output to include terms that are within those ideological paradigms, providing

an opportunity to create alternative scenarios of central bank responses. If such a model were to have these capabilities and were used to assist institutional policy work, then institutions will need to implement rules around the use of these outputs. For central banks, this would result in additional oversight to ensure ethical guidance is produced from these models to maintain transparency and accountability for policy actions.

A downside to the complexity and the size of LLMs, is that models trained to generate policy statements would largely be dependent on a limited number of methods to explain its output. The broad classes of techniques available for model interpretation use challenge sets, adversarial examples, and prediction explainability methods to understand model connections as explored in Danilevsky et al. (2020). However, many of these techniques may not be able to capture the full depth of the most recently published models that have placed restrictions on model code access and are pretrained on datasets not available to the public. These techniques also face another issue of being English dominant, so if models are exposed to non-English datasets it would be difficult to assess them using the most common NLP benchmark datasets and tasks. It is recommended that human oversight be retained in some capacity to ensure that outputs used are aligned with ethical guidelines and government mandates.

As shown in Table 11.4, GPT-1, GPT-2, and BERT models are all thoroughly open-source projects with little accessibility barriers besides technical knowledge. The social cost of having them available allows for more free experimentation and application of these models in both helpful and harmful contexts. For instance, an institution that is interested in training a model to produce Federal Reserve statements for research and a group of malicious actors wanting the same model to spread misinformation would have equal access to reach their goals. In addition to assuring models its utilizing in policy processes, an institution would also need to consider how to handle outside actors, using such a model to impersonate leading policy figures and spread misinformation.

During and after the release of GPT-1, misuse of this model was not apparent and not many safeguards were provided to prevent a user from fine-tuning the model for a negative goal such as spreading fake news. How-

ever, during the development and release of GPT-2, concern about misuse of these models was addressed more directly. OpenAI selected four universities to study different biases and vulnerabilities of the model as iterated in their report focusing on the social impacts of GPT-2 (Solaiman et al., 2019). It was found that LLMs are strongly influenced by the data given to them and are capable of creating believable misinformation. The following year, with the release of GPT-3, a similar thorough report was not released, but was more of a discussion summary of open research questions regarding the model. The report points to an urgent need to discuss model accessibility, inherent model bias, and deliberate harmful use to spread misinformation (Tamkin et al., 2021).

One of the most important aspects of these models that institutions should consider starts at the beginning of an LLM-based project: choosing which pre-trained LLM model should be used. This may be perhaps the most important consideration since picking a pre-made model means also accepting or planning to mitigate any inherent bias. AI central banker could be biased towards a certain monetary policy framework even before the model is fine-tuned based solely on the training data.

11.3.5 Data assurance of LLMs

The foundation for advanced LLM models is the dataset they are being trained on. However, an important study in 2016 raised the issue of gender stereotypes bias appearing in word embeddings after a model was trained on Google news articles (Bolukbasi et al., 2016). The effectiveness of a model can be benchmarked by the data that it is exposed to both in the pre-training process and the fine-tuning process, though datasets involved in either stage of the training process inherently carry some form of bias, and it can be difficult at times to detect the influence it has on model outputs. An obvious example of the impact of data on models and how it influences the functionality of AI is MIT's Norman AI project.

The 2018 Norman AI project was developed by MIT to show how data can significantly impact a machine learning model, similar to the ideological experiments mentioned earlier by Middlebury University. The model was trained on a compilation of dark disturbing content from Reddit, and

was then tasked to provide a textual description of images (Yanardag et al., 2018). The outcome of the experiment was a clinically diagnosed psychopathic AI. Norman was fed a challenge dataset of Rorschach inkblots, and its outputs were compared to another model that was pre-trained on more standard data instead of the dark Reddit dataset. The results showed quite disturbing outputs from Norman that were drastically different from the other model. The MIT project highlights an extreme output created from the data it was exposed to. However, bias can be more subtle in datasets and has remained unfiltered in outputs of influential LLMs. Many studies like the GPT reports on risk and bias focused on how these models could be fine-tuned to have negative bias or be used for harmful actions. The first opportunity for bias in a model derives from the pre-training datasets, as they have been found to contain harmful bias trends that have passed into the foundation for many projects. It is critically important to address and assure the caveats of pre-training datasets used for LLMs, before and after they are fine-tuned.

The pre-training dataset for GPT-1 was BookCorpus, a dataset of over 20,000 unpublished books. During the development of the GPT-1 project, BookCorpus was considered a common textual dataset to work with and has been used to train around thirty important language models, including BERT. In a recent study, it was found that BookCorpus has concerning issues, such as containing duplicate books, copy-right violations, and an unbalanced list of genres (Bandy and Vincent, 2021). A large number of books from the dataset have a disclaimer listed on its cover, making public use of this dataset a violation of their licensing rights.[11] This discovery is a significant issue if the dataset is used for active policy work outside of the research sphere because of potential legal implications. Other issues included only containing 7000 unique books out of the claimed 22,000 books and drawing too many adult romance genre books compared to other topics. The repetitive texts paired with the genre skew create avenues for potential inherent gender bias in text generation for models that use BookCorpus. This would be most problematic in zero-shot setting applications that do not have any

[11] The disclaimer that was found was as follows: "[this book] may not be distributed to others for commercial and non-commercial purposes".

form of bias mitigation, but can also be an issue even if models have some form of assurance.

The issues raised surrounding BookCorpus shows the need for structured documentation surrounding textual datasets and the potential bias in many currently available models. Models such as FinBERT, which are used for economic analysis, can carry bias from their pre-trained state to their fine-tuned model output if the fine-tuned dataset also contains connections to the same bias paradigms. When a model is fine-tuned or built upon, researchers will typically freeze earlier layers of a neural network, while adding in the new later layer that will accommodate the training of more nuanced relationships. This practice carries long-term model functionality and trends through the network, consequently percolating bias into the later layers. On the macroeconomic level, if bias pertaining to certain hawkish or dovish policy actions is inherent in the fine-tuning process, then text generated by these models for central bankers will provide non-neutral language when recommending policy action. To intentionally perpetuate this bias, an analyst could simply compile a series of monetary policy statements from a single individual or central bank that has a history of preferring one policy method or the other and use it as the only fine-tuning dataset.

Even with the expanded set of diverse datasets available today, it remains the case that there is not a perfectly unbiased large dataset. It is still difficult to create a balanced dataset for specific content because of bias in historical trends and data availability. While sometimes the greatest limiting factor for mitigating bias in datasets is time, using smaller balanced datasets can render a model effective only in a small range of tasks. Thus even if smaller datasets are easier to manage, it is not an option to scale dataset size downwards either. As LLMs become more commonly used, there will continue to be a tradeoff between providing a sufficient volume of datasets and effectively mitigating potential bias in a focused dataset. Models will continue to require the creation of large datasets for now; more techniques will need to be developed to identify trends that connect with harmful outcomes. Without data assurance for these large datasets, a model could turn out to be a more elusive version of Norman.

11.3.6 LLM transparency

Many models are trained on private datasets that contain personally identifiable information (PII). Typically, these datasets are medical information datasets that a company acquires through a purchasing agreement and is legally bound to follow strict guidelines regarding its use. One of the main restrictions being that the information in these datasets is not allowed to be shared, distributed, or accessible to the public. Recently a study in collaboration with Google, Apple, OpenAI, and a few top universities, found that, aside from bias issues with pre-training datasets, there is a significant data privacy challenge with LLMs as well (Carlini et al., 2021). The goal of this project was to illustrate how difficult it would be for an individual to perform a data extraction by leveraging the model's actions alone, and is a caution for potential data breaches occurring in the future. The project describes the creation of a successful data extraction process of PII from GPT-2's pre-training dataset (WebText), by exploiting the model with targeted questions. Breaches such as these can cause harm to individuals and is a security concern related to the neural network information retrieval process within LLMs.

If a central bank is using these models to draft policy or regulation for a specific community or sector, it would be important for the policymaker and analyst to secure/discard outputs that leak sensitive information. Public transparency when using these models in both the private and public sector is varied across the field. There aren't clear guidelines on best practices, but there is a call for more rules to be implemented regarding data privacy overall. Some companies have responded by choosing to hide pre-training datasets from the public, such as Google's mysterious large dataset it uses to train models, while others continue to publish the datasets they made along with their models. It is difficult to determine the best measure of data privacy for both pre-training and fine-tuning datasets.

Since the release of these analysis reports and community concerns regarding the use of AI, most of the guardrails for these models are being mentored and created by the model creators themselves. This can be seen in OpenAI's choice to forgo making the GPT-3 project fully open-source, just as it has done with its previous models. As discussed in several points pre-

viously, OpenAI has dedicated a significant number of resources to understanding the risks associated with its GPT project. However, the company has also admitted that it doesn't have the resources to safeguard against professional orchestrated attacks. They have called for government institutions to help protect against such instances from occurring. At the same time, BERT has continued to be open-source despite Google's incorporation of it into its search engine. The accessibility of the code behind the models remains varied, with some companies turning away from having code repositories being completely open-source. So far, the AI community has decided that allowing LLMs and the majority of large datasets created to train LLMs is beneficial to encourage transparency. Nonetheless, in 2020 and 2021, many governments have drafted policy guidance to begin processing guardrails and transparency rules that are more than just recommendations. The greatest challenge will be for these rules to strike a proper balance that will protect against the misuse of these models in economics and finance, while allowing value to be delivered from these models.

The next stage for the GPT project and other advanced language models is unknown, but it is speculated that LLMs would soon be able to learn on just a few examples, similar to how humans learn. The pathway to artificial generalized intelligence (AGI) has become more of a reality with projects like GPT, but with an unclear path forward. LLMs will continue to become larger, more generalized, and harder to understand. Any economic institution will need to ultimately decide what characteristics of a model would be best to use. Whether it should be one that is more open-source, one from a private company, one from a non-profit, one made available only with an API, or one that would have to be purchased, etc. In conclusion, it is important for institutions to develop and use tools to explain model outputs as these models become larger and more complex. It is inevitable that they will take part in the policy process at some point, and it is imperative that institutions are capable of understanding and utilizing them.

11.3.7 Association rules mining

Association rules mining (ARM) is an application of decision rules, which involves making conclusions using common patterns among sets of data to

identify strong associations or "rules." The most common use of association rules mining is to find patterns in transaction data to discover which items are frequently purchased together. Agrawal et al. (1993) was the first to use transaction data. Decision rules derives from decision theory, which has its roots in operations research from WWII to support decision-making using empirical results.

ARM outputs are a series of "if-then" rules, showing co-occurrences between item sets X (antecedent) and item sets Y (consequent) out of N transactions. One can also use regression syntax: X as the left-hand side (LHS) variables and Y the right-hand side (RHS) variables. This means that when items X appear in the data, item(s) Y are likely to be found, too. An example rule for grocery transactions is {beer} $\Rightarrow$ {diapers}, where beer is the antecedent and diapers are the consequent. This would mean that diapers are often purchased when beer is purchased. Note that correlation does not equal causation. How strong is a given association X $\Rightarrow$ Y? There are several metrics to use to understand and evaluate associations:

1. $Support(X) = \frac{\#\ instances\ of\ X}{\#\ N\ transactions}$
2. $Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$
3. $Lift(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) * Support(Y)}$

Support is the ratio of how many of the N transactions include items X. This measure can be used to set limits on how common or uncommon items X should appear for rules mining. The confidence measure is the ratio of how consistently Y appears when X also appears, which can be thought of as the conditional probability of Y happening given X. Finally, lift is the conditional change on the probability of Y appearing when X also appears. When lift is 1 then the two item sets are independent of one another. If lift > 1, then X in a transaction means that Y is more likely to appear. If lift < 1 then X's appearance makes Y less likely. Another way to think about these relationships is that X and Y could be compliment (lift > 1), such as ice cream and pie, or substitutes (lift < 1), such as coconut milk and almond milk.

Using ARM for international trade has been studied by Batarseh et al. (2021b) to evaluate which commodities are frequently traded together. Rather than a shopper buying items from a grocery store, a "transaction" is a

country A going to another country B to choose from a variety of commodities in a specific year. Each category of commodity such as soybeans, electronics, and others are organized into the harmonized system (HS) codes, which are used at high granularity for tariff actions. For research purposes, higher-level aggregations have been considered.[12] Data from the WTO was used to evaluate how commodities are strong consequents or antecedents within bilateral or multi-lateral relationships. In Batarseh et al. (2020), the trade relationship between China and Australia in the aftermath of COVID-19 was studied by analyzing the volume of bilateral association rules that meet a given threshold. This method meant that the authors established equal weighting of rules, digging for the common threads among millions of rules. Harnessing ARM for understanding commodity-level trade data has helped answer questions about trade relationships that traditional economic methods would struggle to discern.

11.3.8 Graph neural networks

Supply chains have grown in complexity due to global economic integration. Consumers in country B rely on raw materials from country A, refinement in their own country B, and manufacturing in country C. AI has contributed to the economic study of trade through early studies into applications of advanced modeling techniques, such as graph neural networks. These trade flows are best described as networks, as demonstrated by the gravity equation of trade in Anderson (1979). The study of network-based (or graph-based) learning has been harnessed to solve problems in molecular dynamics (Duvenaud et al., 2015), traffic prediction (Chen et al., 2020), and stock market prediction (Li et al., 2020). These types of models are called graph neural networks (GNNs).

GNNs are powerful at solving problems where there is a ripple effect from one entity to another. In the case of traffic prediction, an accident at one intersection is likely to cause additional congestion on surrounding roads. This is an example of a node-level prediction task, where all the nodes are being predicted simultaneously. Communication between nodes is referred

[12] Read more about HS codes at the International Trade Administration - https://www.trade.gov/harmonized-system-hs-codes.

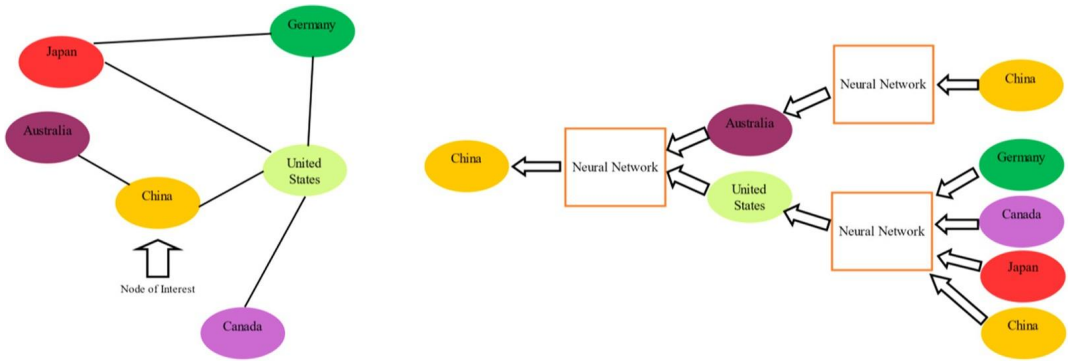


FIGURE 11.6 Diagram showing how GNN prediction is made for node classification by taking other nodes connected by edges.

to as message passing because surrounding nodes are contributing to the information set for prediction at a single node. These neighborhood effects can be adjusted for the degrees of separation to include in the prediction at a particular node (Defferrard et al., 2017). The edges in a GNN are critical to proper flow of information as stronger edges have greater effect than weaker edges.

11.3.8.1 GNNs for international trade

GNNs have been harnessed for several international trade applications in the last few years. Panford-Quainoo et al. (2020) were the first to show GNN modeling of bilateral trade relationships. Using countries as the nodes and bilateral trade as the edges of the network. See Fig. 11.6 diagram of how the GNN is using information from surrounding countries to predict data on China. That framework was able to perform several tasks, including predicting trade links and predicting country income brackets. This model was implemented on a single time period, thus the choice of data could have a substantial impact on the final result.

Expanding on the single period GNN, Monken et al. (2021) implements a time-varying GNN structure for international trade. Combining the elements of a recurrent neural network structure, called long short term memory (LSTM), with GNNs helped the model overcome the challenge of the edges of the network changing significantly over time. Traditional GNNs struggle to formulate effective graph representation when edges change

significantly. The node prediction task being tested was the trade unit value (TUV) of soybeans, using previous volumes and prices of soybeans among trading partners around the world. As a stable food commodity, the choice of data meant that there would be substantial trade among many countries. Data assurance was conducted by identifying any issues with underlying values that were outliers in the distribution as well as large enough volumes to have a real-world trade impact. The training method of this GNN involved combined gradient across all countries such that all countries' predictions contributed equally towards the loss function. This meant that the outcome needed to utilize similar units across all countries. Assuring the model output involved checking error rates across different cross sections of the data, eliminating any possibility for Simpson's paradox or vastly varying model performance across countries. Another aspect of AI assurance implemented was applying counterfactual scenarios, allowing the complexity of the GNN to be empirically tested by perturbing the input data. By altering single edges or nodes, multiple "what-if" scenarios were tested. Further AI assurance using explainability techniques will make GNNs even more useful for economic policymakers.

11.3.8.2 Explainability methods for GNNs

Various interpretability methods for GNNs are discussed in the review paper by Yuan et al. (2020), which includes four broad categories as summarized in Table 11.5. All have the goal of making GNNs more interpretable, thus more easily assured, evaluated, and audited by economic policymakers. Some of these methods build subgraphs to find the most salient connections and features within the larger GNN, while others consider the local environment of the prediction environment for interpretability.

Each of these explanation methods requires background research to implement effectively. Frequently, the code implementation lags months or years behind the paper describing the theoretical method. The most developed explainability method is GNNExplainer, which is available as part of the PyTorch Geometric package. The other tools often have a GitHub repo with someone's code development. As GNNs become more common in economics, more explainability techniques need to be developed in ma-

Table 11.5 Summary of explainability methods for GNNs.

Category Description	Example Tool	Paper
Probe the gradients from model training to determine the most important features	CAM (Class Activation Mapping)	Pope et al. (2019)
Perturb the input data by introducing masks against certain nodes, edges, or node features to find which masks change the output the most	GNNExplainer	Ying et al. (2019)
Construct a localized model containing immediately surrounding features to build an explainable model	GraphLIME	Huang et al. (2020)
Decompose the final prediction for the GNN by stepping back a layer in the network and testing the impact from the input features	GNN-LRP	Schnake et al. (2020)

jor GNN package environments, such as PyTorch Geometric or Tensorflow Geometric.

11.4 Conclusion

This chapter has outlined the current AI methods that are and will continue to reshape the economics domain. AI has already made advances in three important ways: a) improving forecasting capabilities using flexible machine learning models, b) evaluating and predicting policymaking statements with large language models and natural language processing, and c) evaluating trade relationships using graph neural networks and data mining. AI methods, while often more powerful than traditional econometric techniques, require AI assurance to be safely, equitably, and ethically implemented. The landscape of AI assurance in economics has been discussed and three explainability techniques used in economics were covered: LIME, SHAP, and PDP. Economic policymaking institutions, however, must be more accountable than private entities when employing AI models for tasks that can fundamentally change the lives of people.

Acknowledgments

We thank Ricardo Correa for his helpful feedback and comments. We also thank Jennifer Massey and Macro Cagetti for their professional guidance and ongoing support for this research.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 207–216. http://doi.acm.org/10.1145/170035.170072.

Anderson, J.E., 1979. A theoretical foundation for the gravity equation. American Economic Review, American Economic Association 69 (1), 106–116.

Araci, D., 2019. FinBERT: financial sentiment analysis with pre-trained language models. arXiv:1908.10063. https://arxiv.org/abs/1908.10063.

Bandy, J., Vincent, N., 2021. Addressing "documentation debt" in machine learning research: a retrospective datasheet for BookCorpus. arXiv:2105.05241. https://arxiv.org/abs/2105.05241.

Batarseh, F.A., Freeman, L., Huang, C.H., 2021a. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 1–30. https://doi.org/10.1186/s40537-021-00445-7.

Batarseh, F.A., Gopinath, M., Monken, A., Gu, Z., 2021b. Public policymaking for international agricultural trade using association rules and ensemble machine learning. Machine Learning with Applications 5, 100046.

Batarseh, F.A., Munisamy, G., Monken, A., 2020. Artificial Intelligence Methods for Evaluating Global Trade Flows. Board of Governors of the Federal Reserve System International Finance Discussion Papers. p. 1296.

Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. arXiv:1607.06520. https://arxiv.org/abs/1607.06520.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pd.

Brusseau, J., 2021. AI human impact: toward a model for ethical investing in AI-intensive companies. Journal of Sustainable Finance & Investment, 1–28. https://doi.org/10.2139/ssrn.3648545.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., Raffel, C., 2021. Extracting training data from large language models. arXiv:2012.07805. https://arxiv.org/abs/2012.07805.

Carrillo, A., Cantú, L.F., Noriega, A., 2021. Individual explanations in machine learning models: a survey for practitioners. arXiv:2104.04144. https://arxiv.org/abs/2104.04144.

Cesaro, J., Gagliardi Cozman, F., 2020. Measuring unfairness through game-theoretic interpretability. In: Cellier, P., Driessens, K. (Eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. In: Communications in Computer and Information Science, vol. 1167. Springer, Cham.

Chen, K., Chen, F., Lai, B., Jin, Z., Liu, Y., Li, K., Wei, L., Wang, P., Tang, Y., Huang, J., Hua, X.S., 2020. Dynamics patio-temporal graph-based cnns for traffic prediction. arXiv:1812.02019. https://arxiv.org/abs/1812.02019.

Cook, T., Gupton, G., Modig, Z., Palmer, N., 2021. Explaining Machine Learning by Bootstrapping Partial Dependence Functions and Shapley Values. KC Fed Research Working Papers. RWP 21-12.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P., 2020. A survey of the state of explainable ai for natural language processing. arXiv:2010.00711. https://arxiv.org/abs/2010.00711.

Defferrard, M., Bresson, X., Vandergheynst, P., 2017. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). https://dl.acm.org/doi/10.5555/3157382.3157527.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1.

Duvenaud, D.K., Maclaurin, D., Aguileraiparraguirre, J., Gomezbombarelli, R., Hirzel, T.D., Aspuruguzik, A., Adams, R.P., 2015. Convolutional networks on graphs for learning molecular fingerprints. In: Proceedings of NIPS, pp. 2224–2232.

Feldman, T., Peake, A., 2021. End-to-end bias mitigation: removing gender bias in deep learning. arXiv:2104.02532. https://arxiv.org/abs/2104.02532.

Freitas, A.A., 2014. Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter 15 (1), 1–10. https://doi.org/10.1145/2594473.2594475.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Geraats, P.M., 2002. Transparency of Monetary Policy: Does the Institutional Framework Matter? University of Cambridge. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.199.8707&rep=rep1&type=pdf.

Giudici, P., Raffinetti, E., 2020. Shapley-Lorenz Decompositions in EXplainable Artificial Intelligence. SSRN working paper.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, Bengio, Y., 2014. Generative adversarial networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'14). https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Gramespacher, T., Posth, J.A., 2021. Employing explainable AI to optimize the return target function of a loan portfolio. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.693022.
Hammer, C., Kostroch, M.D.C., Quiros, M.G., 2017. Big Data: Potential, Challenges and Statistical Implications. International Monetary Fund Staff Discussion Notes.
Harrison, D., Rubinfeld, D., 1978. Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management 5 (1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2.
Hassan, T.A., Hollander, S., van Lent, L., Tahoun, A., 2019. Firm-level political risk: measurement and effects. The Quarterly Journal of Economics 134 (4), 2135–2202.
Hassan, T.A., Hollander, S., van Lent, L., Tahoun, A., 2020. Firm-Level Exposure to Epidemic Diseases: Covid-19, SARS, and H1N1. NBER Working Papers 26971. National Bureau of Economic Research, Inc.
Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., Chang, Y., 2020. GraphLIME: local interpretable model ex-planations for graph neural networks. arXiv:2001.06216. https://arxiv.org/abs/2001.06216.
Hurlin, C., Pérignon, C., Saurin, S., 2021. The Fairness of Credit Scoring Models. SSRN Working paper.
Kamiran, F., Calders, T., 2017. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1–33. https://doi.org/10.1007/s10115-011-0463-8.
Khalid, Balar, Rachid, Chaabita, 2019. Big data in economic analysis: advantages and challenges. International Journal of Social Science and Economic Research 4 (7), 5196–5204.
Li, Wei, Bao, Ruijan B., Harimoto, K., Chen, Deli, Xu, J., Su, Q., 2020. Modeling the stock relation with graph network for overnight stock movement prediction. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence Special Track on AI in FinTech, pp. 4541–4547.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
Marwala, T., Hurwitz, E., 2017. Artificial intelligence and economic theories. arXiv: 1703.06597. https://arxiv.org/abs/1703.06597.
McGuffie, K., Newhouse, A., 2020. The radicalization risks of GPT-3 and advanced neural language models. arXiv:2009.06807. https://arxiv.org/abs/2009.06807.
Misheva, B.H., Osterrieder, J., Hirsa, A., Kulkarni, O., Lin, S.F., 2021. Explainable AI in credit risk management. arXiv:2103.00949. https://arxiv.org/abs/2103.00949.
Moloi, T., Marwala, T., 2020. Introduction to artificial intelligence in economics and finance theories. Artificial Intelligence in Economics and Finance Theories, 1–12.
Monken, A., Haberkorn, F., Gopinath, M., Freeman, L., Batarseh, F., 2021. Graph neural networks for modeling causality in international trade. In: The International FLAIRS Conference Proceedings, p. 34.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. The Journal of Economic Perspectives 31 (2), 87–106.
Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116 (44), 22071–22080.

Navarro, C., Kanellos, G., Gottron, T., 2021. Desiderata for explainable AI in statistical production systems of the European Central Bank. arXiv:2107.08045. https://arxiv.org/abs/2107.08045.

Nesvijevskaia, A., Ouillade, S., Guilmin, P., Zucker, J.D., 2021. The accuracy versus interpretability trade-off in fraud detection model. Data & Policy 3, E12.

Nori, H., Jenkins, S., Koch, P., Caruana, R., 2019. Interpretml: a unified framework for machine learning interpretability. arXiv:1909.09223. https://arxiv.org/abs/1909.09223.

Ohana, J.J., Ohana, S., Benhamou, E., Saltiel, D., Guez, B., 2021. Explainable AI (XAI) models applied to the multi-agent environment of financial markets. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 189–207.

Pace, R.K., Barry, R., 1997. Sparse spatial autoregressions. Statistics & Probability Letters 33 (3), 291–297. https://doi.org/10.1016/S0167-7152(96)00140-X.

Panford-Quainoo, K., Bose, J., Defferrard, M., 2020. Bilateral trade modeling with graph neural networks.

Parkes, D.C., Wellman, M.P., 2015. Economic reasoning and artificial intelligence. Science 349 (6245), 267–272.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q., 2017. On fairness and calibration. In: Proceedings of the 31th International Conference on Neural Information Processing Systems (NIPS'17). https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf.

Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E., Hoffmann, H., 2019. Explainability methods for graph convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10764–10773.

Preuss, B., 2021. Contemporary Approaches for AI Governance in Financial Institutions. Copenhagen Business School.

Radford, A., Narasimhan, K., 2018. Improving Language Understanding by Generative Pre-Training. OpenAI. https://openai.com/blog/language-unsupervised/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners. OpenAI. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rei, M., 2017. Semi-supervised multitask learning for sequence labeling. arXiv:1704.07156. https://arxiv.org/abs/1704.07156.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv:1606.05386. https://arxiv.org/abs/1606.05386.

Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M.A., Bravo, C., 2021. Super-app behavioral patterns in credit risk models: financial, statistical and regulatory implications. Expert Systems with Applications 169, 114486.

Rosset, C., 2020. Turing-NLG: a 17-billion-parameter language model by Microsoft. Microsoft Research Blog 2, 13.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schutt, K.T., Muller, K.R., Montavon, G., 2020. Higher-order explanations of graph neural networks via relevant walks. arXiv:2006.03589. https://arxiv.org/abs/2006.03589.

Schwab, K., 2016. The fourth industrial revolution. What it means and how to respond? https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.
Severino, M.K., Peng, Y., 2021. Machine learning algorithms for fraud prediction in property insurance: empirical evidence using real-world microdata. Machine Learning with Applications 100074.
Shapley, L.S., 1953. A value for n-person games. In: Contributions to the Theory of Games, vol. 2.28, pp. 307–317.
Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B., 2019. Megatron-lm: training multi-billion parameter language models using model parallelism. arXiv:1909.08053. https://arxiv.org/abs/1909.08053.
Simpson, Edward H., 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B, Methodological 13 (2), 238–241.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., Wang, J., 2019. Release strategies and the social impacts of language models. arXiv:1908.09203. https://arxiv.org/abs/1908.09203.
Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., Liu, W., Wu, Z., Gong, W., Liang, J., Shang, Z., Sun, P., Liu, W., Ouyang, X., Yu, D., Tian, H., Wu, H., Wang, H., 2021. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. arXiv:2107.02137. https://arxiv.org/abs/2107.02137.
Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., Wu, H., 2019a. Ernie: enhanced representation through knowledge integration. arXiv:1904.09223. https://arxiv.org/abs/1904.09223.
Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., Wang, H., 2019b. Ernie 2.0: a continual pre-training framework for language understanding. arXiv:1907.12412. https://arxiv.org/abs/1907.12412.
Taborda, B., Almeida, A., Carlos Dias, J., Batista, F., Ribeiro, R., 2021. Stock Market Tweets Data. IEEE Dataport.
Tamkin, A., Brundage, M., Clark, J., Ganguli, D., 2021. Understanding the capabilities, limitations, and societal impact of large language models. arXiv:2102.02503. https://arxiv.org/abs/2102.02503.
Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard Journal of Law & Technology 31, 841.
Yanardag, P., Cebrian, M., Rahwan, I., 2018. Norman AI Project. MIT. http://norman-ai.mit.edu/.
Yang, Y., Zheng, Z., Weinan, E., 2020a. Interpretable neural networks for panel data analysis in economics. arXiv:2010.05311. https://arxiv.org/abs/2010.05311.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V., 2020b. Xlnet: generalized autoregressive pre-training for language understanding. arXiv:1906.08237. https://arxiv.org/abs/1906.08237.
Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. GNNExplainer: generating explanations for graph neural networks. arXiv:1903.03894. https://arxiv.org/abs/1903.03894.

Yuan, H., Yu, H., Gui, S., Ji, S., 2020. Explainability in graph neural networks: a taxonomic survey. arXiv:2012.15445. https://doi.org/10.48550/arXiv.2012.15445.
Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, New York, NY, USA, pp. 335–340.
Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D.C., Socher, R., 2020. The AI economist: improving equality and productivity with AI-driven tax policies. arXiv:2004.13332. https://arxiv.org/abs/2004.13332.

This page intentionally left blank

12

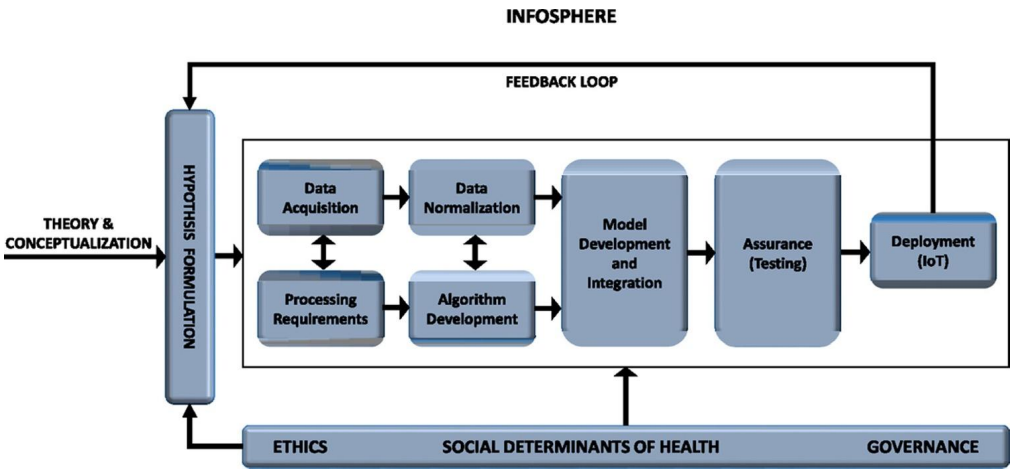
Panopticon implications of ethical AI: equity, disparity, and inequality in healthcare

Erik W. Kuiler and Connie L. McNeely
George Mason University, Fairfax, VA, United States

*A first-rate pilot (*cybernetes*) or physician feels the difference between possibilities and impossibilities in his art and attempts the one and lets the others go; and then, too, if it happens that he does trip, he is equal to correcting his error.*
—Plato's *Republic*, 360e-361a

...computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology. I use the phrase "computer technology" because I take the subject matter of the field broadly to include computers and associated technology. For instance, I include concerns about software as well as hardware and concerns about networks connecting computers as well as computers themselves.
—Moore, 1985

Graphical abstract



AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00026-3

Abstract

Addressing issues in the healthcare domain as principal illustrations, instrumentalities of Artificial Intelligence (AI) are explored at the nexus of ubiquity and control reflected in the Internet of Things (IoT) and in various governance and policy features. Roles of ethics and ontological determinations are examined with particular attention to the effects of AI-enabled systems and policies on selected groups and related health disparities and biases. The growth of predictive data analytics and the simultaneous growth in the availability of interoperable AI-enabled devices offer opportunities to mitigate healthcare disparities currently endemic in indigent, underrepresented, and underserved communities. However, use of these devices can exacerbate inequities as well as ameliorate them. By themselves, human and AI agents operating in an IoT-sustained infosphere cannot solve problems related to the multiple forms of marginalization that affect the health and wellbeing of particular communities. They require actively engaged and empathetic governance to address complex socioeconomic issues and solve complex accessibility and distribution problems. In collaborative environments, where healthcare provision depends on the synergy of human and AI actors, regulatory models must not only address the ethical conduct of medical practitioners, but also the professional conduct of informaticists, software developers, and device vendors.

Keywords

Artificial Intelligence, ethics, healthcare, infosphere

Highlights

- By themselves, human and AI agents operating in an Internet of Things (IoT)-sustained infosphere cannot solve the multiple forms of marginalization that affect the health and wellbeing of particular communities.
- In collaborative environments, where healthcare provision depends on the synergy of human and AI actors and systems, regulatory models must not only address the ethical conduct of medical practitioners, but also the professional conduct of informaticists, software developers, and device vendors.
- Human and AI agents require actively engaged and empathetic governance to address complex socioeconomic issues and solve complex accessibility and distribution problems.

- Human and AI agents operating in an IoT-sustained infosphere require actively engaged and empathetic governance to address complex healthcare issues and solve related socioeconomic, accessibility, and distribution problems. In the broader infosphere and collaborative environments, healthcare provision depends on the synergy of human and AI actors within regulatory models.

12.1 Introduction

The advent of the Internet and the availability of relatively inexpensive high-speed computers have facilitated the integration of human and machine agents in a global network of information-based digital capabilities used to deliver products and services on a world-wide scale. This situation, while directly or indirectly touching virtually all aspects of life in one way or another, has been particularly transformative in medical practice and healthcare provision. Over the last 25 years, the medical and healthcare domain has witnessed an extraordinary growth in the artificial intelligence (AI) supported, big data-driven practice of medicine. This growth has been evidenced in the introduction of electronic health records (EHRs), personal health records (PHRs), digital imaging, digitized procedures, increasing sophistication in laboratory test formulation, real-time availability of sensor data, and the introduction of genomics-related projects, among other things. AI support in healthcare frequently incorporates techniques such as machine learning (ML) and natural language processing in, for example, diagnostic expert systems, clinical decision support systems, robotics, and process automation (Kuiler and McNeely, 2018, 2020).

AI is concerned with "the computational understanding of what is commonly called intelligent behavior, and with the creation of artifacts that exhibit such behavior" (Ramesh et al., 2004). The term "augmented intelligence" also has served as a conceptualization of AI, focusing on its assistive role and emphasizing that its design "enhances human intelligence rather than replaces it" (AMA, 2021). AI has its provenance in diverse disciplines—mathematics, computational sciences, psychology, biology, semiotics, linguistics, and philosophy—and reflects the use of inductive, deductive, and abductive logic systems. (Peirce, 1958, 1997; Kuiler and McNeely, 2020)

As an area of study, AI reflects different perspectives—cultural, scientific, and technological—and depends on a range of methodological approaches to ensure its efficacy. AI-facilitated data analytics typically are posited as means for enhancing wellbeing and, with algorithm-based procedures, can help address complex socioeconomic issues, reduce administrative burdens, and solve multifaceted distribution problems, as suggested by current pandemic-focused logistics efforts. However, AI-enabled systems also can reflect or lead to problems and challenges to existence and life chances, and have been invoked in terms of, for example, perpetuating discrimination (intentional or unintentional) and abrogating rights (e.g., invasions of privacy). It is in this sense that questions arise in relation to AI assurance regarding ethics in the design, development, implementation, and operationalization of AI-enabled instruments and policies, and motivating the analytical direction of the research presented here.

While considering affective issues relative to ethical AI and assurance more generally, the medical and healthcare domain is engaged here as a principal illustration and analytical focus in light of its centrality on public agenda and prominence in AI-driven practice. Since the mid-1990s, a debate has taken place within the medical informatics community on how to develop and integrate AI agents with the sound, ethical practice of medicine and healthcare provision (Lloyd, 1985; Coiera, 1996; Kulikowski, 1996). In this regard, the roles of ethics and ontological determinants are examined with particular attention to medical and healthcare issues among indigent, underserved, and underrepresented communities. Such populations may be especially affected by AI-enabled policies and their operationalization reflected in, for example, geospatial divides, digital and knowledge divides, and other eudaemonic disparities, whose effects transcend technological inequalities and inequities. Addressing disparities in healthcare access and provision that afflict disadvantaged and disenfranchised communities in many parts of the world, a conceptual framework and analysis are provided in reference to ethical instrumentalities of AI explored at the nexus of digital ubiquity, especially based on the omnipresence of the Internet of Things (IoT) and control, as delineated in issues of governance and regulation.

12.2 Ontological perspectives

Ontologies encapsulate the intellectual histories of epistemic communities and support the development of dynamic heuristic instruments that sustain science and society. Noting an Aristotelean tradition, the universe has been conceptualized as comprising entities that consist of form and matter, where form is, or contains, information (cf. *Metaphysics*, Book VII.3; Bynum, 2010). Moreover, human beings are distinguished from other forms of life, such as plants or animals, by the human abilities for theoretical and practical reasoning (cf. *On the Soul*, Book III.3). Thus, information-processing capabilities are encapsulated in physical bodies (Bynum, 2000, 2010).

In more modern terms, "information is not just an abstract concept... It is a concrete property of matter and energy that is quantifiable and measurable. It is every bit as real as the weight of a chunk of lead or the energy stored in an atomic warhead, and just like mass and energy, information is subject to a set of physical laws that dictate how it can be manipulated, transferred, duplicated, erased or destroyed. And everything in the universe must obey the laws of information because everything in the universe is shaped by the information it contains" (Seife, 2006, p. 2). The concept of cybernetics is relevant to this thinking, adumbrating Aristotelean notions of matter and form, information and reasoning, by positing that it is in the nature of the universe that all entities consist of information encoded in matter-energy (Wiener, 1948, 1954). That is, humans and computerized machines are cybernetic entities; they are dynamic systems with components that communicate internally and externally with the outside world by means of information channels and feedback loops. In this context, society comprises communities of intra- and inter-dependent cybernetic entities determined by information generation and exchange (Wiener, 1954; Bynum, 2010; Boulding, 1956, 1966).

Panopticon effects of AI on the infosphere

As makers, manipulators, and encapsulations of information, human intelligence and AI entities—collectively, cybernetic entities—operate as agents in a universe of information, in an "infosphere" (Boulding, 1970,

1980; Floridi, 2010b). The infosphere is born in the matriculation of artificial and human intelligences and their interactions (Floridi, 2010b; Bynum, 1985). As such, it is central to discussions of information ethics in which information fulfills three interdependent roles: as a resource, as a target, and as a product (Floridi, 2010b).

In the healthcare domain, the interactions of these human and artificial intelligent agents are operationalized via the IoT, resulting in a panopticon effect. Telemedicine, for example, makes healthcare available to many individuals who otherwise would not have ready access to medical expertise or facilities. AI/ML-enabled software as a medical device (SaMD) applications allow medical practitioners to monitor the conditions of their patients at all times, as long as secure information and communications technology (ICT) connectivity can be maintained. SaMD performs on non-medical devices, such as smartphones, smartwatches, laptops, tablets, or other computing platforms. SaMD, for example, allows viewing of magnetic resonance images for diagnostic purposes from a smartphone (FDA, 2021a,b; IMDRF, 2020). Note that software that relies on data received from a medical device but that does not have a medical purpose are not considered SaMD. Rather, software that, for example, encrypts data transmissions from a medical device or software that controls the motor for pumping medication in an infusion pump, MRI, EKG, EHR, and X-ray machines is designated software *in* a medical device (SiMD) (FDA, 2021b). The graphical abstract provides a conceptual framework of the infosphere, highlighting the importance of ethics, sociocultural factors, and governance in the development and deployment of AI agents in IoT environments.

The implementation of assurance regimens instill confidence that AI agents in the infosphere operate within their appropriate ethical, regulatory, and legal frameworks. AI assurance is a technical specialty that complements the independent verification and validation (IV&V) tasks of industry-standard system development life cycles (SDLC). AI assurance processes are designed to test the behavior of algorithms and the integrity and accuracy of the data to which those algorithms are applied. Aspirationally, AI assurance emphasizes that AI systems should be trustworthy and fair by addressing social determinants equitably and without preju-

dice, incorporate relevant legal and ethics codes, and comprise explainable models with regard to data, algorithms, and their interactions (Batarseh et al., 2021). Also, AI systems should be safe and secure. As a specialty domain, AI assurance focuses on the independent validation and verification of AI systems and their platforms, their quality and reliability, their robustness and efficacy, their security and resilience to adversarial attacks, and their privacy protection mechanisms. Furthermore, AI systems should reflect thorough analyses and assessments of risks. It is also in this regard that US Food and Drug Administration (FDA) regulations require that appropriately constituted institutional review boards be formally designated to evaluate and monitor all biomedical research that involves human subjects and to protect their rights and welfare (FDA, 2020).

12.3 Ethics frameworks

In the context of AI development, the purpose and application of codes of ethics are to delineate the moral dimensions of the development, introduction, and conduct of AI agents. They also apply to the systematic computational analytics of structured and unstructured data and their attendant outcomes to guide the conduct of actors engaged in those activities.

Infosphere ethics

The conceptualization of a demiurge *homo poieticus* as the bridge builder between *physis* (*phusis*) and *techne* in the infosphere has been used in the recent literature on macroethics.[1] From a macroethical perspective, human beings, qua *homo poieticus*, are moral agents who have obligations to maintain the infosphere and help it to flourish and increase in value (Floridi, 1999, 2018; Floridi and Sanders, 2004, 2005). These obligations constitute principles focusing on negative entropy (Floridi, 2010a,b, p 92):

[1] Given that *physis* has active as well as passive aspects, it is not clear how this bridge can be created without *nomos*, *episteme*, and *phronesis* (cf. Floridi and Sanders, 2004, 2005). For an earlier discussion of the interdependencies between *techne* and *episteme*, *physis* and *nomos*, in social, communitarian contexts, see Aristotle's *Nicomachean Ethics*, Book V and VI; on *phronesis*, see Book VI). See also Plato's discourse on *physis* and *nomos* in *Gorgias*, in which Callicles argues that it is a matter of justice according to *physis* for the strong to prey on the weak.

- Entropy ought not to be caused in the infosphere.
- Entropy ought to be prevented in the infosphere.
- Entropy ought to be removed from the infosphere.
- The flourishing of informational entities, as well as those of the whole infosphere, ought to be promoted by preserving, cultivating, and enriching their properties.

In the society more generally, the mitigation of entropy is necessary to support an ethical framework; however, it is not sufficient. Moral human agents are also obligated to uphold principles of justice and freedom (Wiener, 1954, pp. 105-106; Bynum, 2010)[2]:

- *Freedom*: the liberty of each human being to develop in his freedom the full measure of the human possibilities embodied in him
- *Equality*: the equality by which what is just for A and B remains just when the positions of A and B are interchanged
- *Good Will*: the good will between man and man that knows no limits short of those of humanity itself
- *Minimum infringement of freedom*: the compulsion the very existence of the community and the state may demand must be exercised in such a way as to produce no unnecessary infringement of freedom.

Healthcare Domain Ethics

An important issue that must be addressed is how to balance the benefits and risks associated with the introduction of AI in the medical practice and healthcare (Rigby, 2019). The benefits of integrating AI in medical practices can be systemic by improving the care of patients and the efficacy of medical procedures. Nevertheless, the need remains to minimize ethical risks of AI implementation. In the healthcare domain, the previously discussed macroethic principles are transformed into deontological ethics that are epistemic and professional in application, focusing on the conduct of medical and ICT professionals as informaticists in their respective spheres of competence. Emphasizing consequentialism, i.e., engaging in actions that cause more good than harm, healthcare ethics apply to the interactions

[2] See also Bynum, 2000; UNESCO, 2019a,b, 2021.

Table 12.1 Foundational Ethics.

Healthcare Practitioners	Informaticists
Beneficence: All persons have a duty to advance the good of others	*Information-privacy and disposition*: All persons have a fundamental right to privacy and to control the data about themselves
Non-malfeasance: All persons have a duty to prevent harm to other persons	*Openness*: The management and disposition of personal data must be disclosed in an appropriate and timely fashion to the subject of those data
Autonomy: All persons have a fundamental right to self-determination	*Access*: The subject of an electronic record has the right of access to that record and to correct the record with respect to its accurateness, completeness, and relevance
Equality and justice: All persons are equal as persons and have a right to be treated accordingly	*Accountability*: Any infringement of the privacy rights of the individual person, and of the right to control over person-relative data, must be justified to the affected person in good time and in an appropriate fashion
Integrity: All persons must fulfill their obligations to the best of their abilities	*Security*: Data that have been legitimately collected about a person should be protected by all reasonable and appropriate measures against loss, degradation, unauthorized use, destruction, access, use, manipulation, or modification
Impossibility: All rights and duties hold subject to the condition that it is possible to meet them under the circumstances that obtain	*Legitimate infringement*: The fundamental right of control over the collection, storage, access, use, and disposition of personal data is conditioned only by the legitimate, appropriate, and relevant data-needs of a free, responsible, and democratic society, and by the equal and competing rights of other persons
Proportionality: All positive features and benefits must be balanced against negative features and risks	*Least intrusive alternative*: Any infringement of the privacy rights of the individual person, and of the individual's right to control over person-relative data may only occur in the least intrusive fashion and with a minimum of interference with the rights of the affected person

between healthcare providers and their patients (AMA, 2020). In their professional conduct, informaticists are also expected to adhere to codes of ethics that reflect universal principles. The guiding ethical principles for healthcare and medical practitioners and informaticists are outlined in Table 12.1 (cf. AMIA, 2013; AMA, 2020; IMIA, 2016; NIH, 2014).

All medical professionals and informaticists are expected to abide by the rules and legal dicta that apply to their epistemes. In the United States (US), for example, professionals who are involved in medical research are also expected to abide by the Federal Policy for the Protection of Human Subjects ("Common Rule") (DHHS, 1991, 2016) and the World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects (WMA, 2018). It is expected that every legitimate medical research project will have been subjected to a rigorous peer review process and will have received the appropriate approvals. In addition, medical protocols, practices, and research endeavors should be evaluated continually for their safety, effectiveness, efficiency, accessibility, and quality (WMA, 2018). Along the same lines, computer scientists are expected to abide by a code of ethics that stipulates that they will not use computers to harm others, to steal intellectual property, and that they should be aware of the social consequences of the systems they are developing (CEI, 2011).[3] Medical device developers are expected to meet technological end regulatory challenges, such as accessibility, interoperability, cybersecurity, data integrity, and data security (IMDRF, 2020).

12.4 Governance in the healthcare domain

Governance is a political process that imposes the structures, norms, and rules that apply to matriculation in society. Aspirationally, governance should incorporate inclusivity, transparency, participation, and responsiveness systemically as foundational principles. Such policy options include binding law, self-regulation (identification and architecture), co-regulation (privacy and ethics), standards and standardization, and *laissez faire* ("do nothing"). In an AI environment, self-governance requires building trust in industry, regulation, and liability mechanisms; the introduction of AI risks mitigation practices across the AI development and implementation life cycle; and the introduction of such practices through widely publicized certification and accreditation regimens (Röösli et al., 2021).

[3] See Moore (1999) for a discussion of just consequentialism in computing. See Tavani (2010) for a discussion of foundationalist debates in computer ethics.

In an IoT environment, where the practice of medicine depends on the collaboration of human and AI actors, regulatory models must address not only the conduct of medical and healthcare providers, but also the professional conduct of informaticists, software developers, and device vendors. Internet connectivity is ubiquitous. On the world stage, set forth in instruments such as the Mauritius Declaration, the result of the International Conference of Data Protection and Privacy Commissioners in Mauritius (EDPS, 2014), principles and recommendations have been formulated to reduce the risks associated with data collected via interconnected devices in the big data ecosystem. Data may only be collected with consent and should be assessed for privacy impacts and data anonymization requirements. The assertion is that individuals own their data and can control them. In this context, self-determination is invoked as an inalienable right of all human beings. (In the US, this right is encapsulated in the Health Insurance Portability and Accountability Act, or HIPAA, regulations). The idea is that data collected in an IoT environment should be "high in quantity, quality, and sensitivity" (EDPS, 2014). Developers and vendors of IoT-connected devices should be transparent and open with regard to the purpose for which the data are collected, how they will be used, how long they will be retained, their retirement, and their disposition. The functional principle is that privacy and rules of ethics should be incorporated by design in IoT innovations, and constructive debate on the technological and ethical implications of the IoT should be pursued and sustained.[4]

Sound governance depends on compliance with societal and epistemic ethics and norms. In AI-IoT environments, this includes explainability and interpretability, not only to increase awareness of algorithm complexities, but also to ensure operational fairness and equity. In addition, it requires an ethical auditing of algorithmic systems to mitigate the risks of social, cultural, and economic inequities (Cath, 2018; Gasser and Almeida, 2017). In the US, the Department of Health and Human Services establishes policies

[4] FDA has proposed a framework for developing machine learning tools in healthcare. The framework provides a useful starting point for discussing the development of ethics-based AI, but does not fully address health disparities in pre- and post-marketing stages of product reviews (Ferryman, 2020).

for the governance of the American healthcare domain. While ICT-enabled technologies provide opportunities to incorporate new functions and features in medical devices, these new technologies also pose direct challenges to existing governance practices: data anonymity and integrity, device interoperability, accessibility, and cybersecurity. The FDA, which in the US regulates SaMD, is also a member of the International Medical Device Regulators Forum (IMDRF), a voluntary group of nation-states that collaborate to harmonize regulatory requirements for medical products that vary from country to country, focusing, for example, on such topoi as unique device identification (UDI), personalized medical devices, standards, adverse event terminology, good regulatory review practices, clinical evaluation, and regulated product submission. The US National Institute of Standards and Technology (NIST) has published standards to address IoT cybersecurity pursuant to the Internet of Things Cybersecurity Improvement Act of 2020 (NIST, 2019, 2020; US Congress, 2020).

12.5 Societal disparities in wellbeing

Governance operates at the nexus of different dimensions—who holds power, who makes decisions, who may participate ("stakeholders"), and how accountability and transparency are instituted—and how these dimensions and their interdependencies are measured and formulated as policies that accord fairness and equity to indigent, underserved, and underrepresented communities to ameliorate their states of wellbeing.

Social determinants of health

Collectively, SaMD, SiMD, telemedicine, ICT-based portals, and other AI agents have far-reaching panopticon effects in addressing the health and wellbeing disparities that, in particular, afflict disadvantaged communities. However, AI agents operating in an IoT environment *per sé* cannot solve the multiple forms of marginalization that may affect the wellbeing of these communities, who are expected to overcome significant, quotidian barriers to wellbeing, usually without general public awareness. These situations reflect complex interactions of various social determinants of health,

Table 12.2 Social Determinants of Health.

Social and Economic Environments	Physical Environment	Cultural Environment	Personal Circumstances
Economic instability: access to employment and unemployment support	*Substandard housing and living conditions*: lack of livable housing, safe neighborhoods	*Intolerance of culture or social norms*: religious intolerance, biased media characterization	*Language barriers*: limited facility and fluency
Societal instability: unequal access the juridical systems, likelihood or presence of conflict	*Geospatial dispersion*: population density per standard geographic area, geographic distance to neighbors	*Social exclusion*: racism and racialized legal status, gender, sexual orientation	*Food insecurity*: irregular access to food sources, limited healthy diet options, limited availability of food banks
Education inaccessibility: unequal access to early childhood education, vocational training, higher education		*Diaspora*: immigration provenance or history, social stigma and discrimination, effects of colonialism, neocolonialism	*Inadequate personal health*: personal or family member addiction, pre-existing medical conditions, including mental health, stress, biology, genetic endowment
Social support unavailability: lack of social safety nets, social capital, social inclusion			*Inadequate family health*: limited or no access to elder care, healthy child development options
Health system inaccessibility: remoteness from healthcare facilities, lack of access to healthcare providers, healthcare unaffordability			*Immobility*: lack of transport, limited personal mobility

as summarized in Table 12.2.[5] Furthermore, sociocultural and economic inequities contribute significantly to premature death and diseases, partic-

[5] Based in part on Henry, 2021; Magnan, 2017; Islam, 2019; Solar and Irwin, 2010.

ularly among vulnerable groups, such as the elderly, disabled, women, and children.

Bias in the healthcare domain

Growing evidence demonstrates the impact of bias on the accessibility, provisioning, and research of healthcare that reflects the social inequalities and inequities inherent in the delivery of healthcare. It is not uncommon, for example, to overlook differential effects and outcomes relative to gender and minority status in clinical research (Kannan et al., 2019). Accordingly, there is an increasing need to analyze and assess deficiencies in the performance of AI models (data and algorithms) in healthcare along the same lines (McCradden et al., 2020). Bias is systemic, long-established, and pervasive and may include, in addition to algorithmic bias, such data biases as historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, population bias, sampling bias, cross-platform behavioral bias, presentation and ranking biases, cause-effect bias (mistaking correlation for cause and effect), observer bias, and so on (Mehrabi et al., 2019).

Arguably, current medical epistemes tend to reflect automation bias, an excessive, *de rigueur* reliance on ICT-based automation (Gianfrancesco et al., 2018; Goddard et al., 2012). For example, it was expected that AI-based automation would provide medical solutions during the COVID-19 pandemic. However, the dissemination of rapidly developed models that reflected and exacerbated cultural, data model, and algorithmic biases may have adversely affected existing health disparities (Röösli et al., 2021). Although efforts have been made to address epistemic cultural and AI-specific technological biases in model building and algorithm development, the challenge remains to prevent adaptive models becoming biased over time. This is the case even for those that incorporate hypothetically fair algorithms, by the creation of feedback loops that reinforce and perpetuate existing biases over time. For example, an algorithm to predict patient mortality or an individual response to treatments could "learn" from existing socioeconomic, racial, and ethnic disparities in care and predict worse treatments for those patients.

In addition, the interactions of AI agents in IoT environments can disseminate implicit and explicit biases among those agents. Medical AI

agents may perpetuate implicit and explicit privilege-based biases that disproportionately benefits individuals from groups who are already privileged over individuals from lesser-privileged groups. Also, even hypothetically fair AI agents can only promote those courses of action that they are intended to promote, which is intentional bias (DeCamp and Lindvall, 2020). As another example, phenotyping using EHRs could benefit from an increasing focus on fidelity, both in the sense of increasing richness, such as measured levels, degree or severity, timing, probability, or conceptual relationships, and in the sense of reducing bias (Hripcsak and Albers, 2018).

Disparities in the healthcare domain

There are sociocultural and socioeconomic issues that cannot be addressed effectively by information and communication technologies alone, increasing the need to address disparities in healthcare access, delivery, and informatics (Veinot et al., 2019). The internet provides the technological foundations for portals, telemedicine and telehealth operations, SaMD, SiMD, and other IoT agents that are the means to address the healthcare inequities that afflict indigent, underserved, and underrepresented communities. However, there are limitations and costs associated with the use of related technologies to address healthcare disparities. Telemedicine, for example, provides not only a means to address socioeconomic and cultural disparities, but also a means to address geospatial proximity, transportation availability, and demographic density issues. Accordingly, telemedicine services, such as portals, are useful for scheduling and coordinating care; for recording personal measurements such as diet, food intake, bodily measurements; and for monitoring home care patients. However, there are also costs associated with such services. For example, there may be security and unreliable connectivity problems. Moreover, not all diagnoses can be done virtually; office visits may still be necessary. Furthermore, healthcare in the US is a market-based commodity. As such, not all telemedicine services are covered by healthcare insurance companies, and both patients and providers must bear the out-of-pocket costs. Also, indigent patients and small medical practices may not have the means necessary to acquire reliable internet connectivity, equipment, and telemedicine products and services. Frequently, there are indirect costs and overhead associated with

telemedicine, and healthcare providers must also bear the administrative costs of the healthcare services that they provide.

Regardless of levels of sophistication, AI-facilitated transaction processing operations, data collection paradigms, and heuristic algorithms *per sé* cannot address the healthcare disparities that afflict indigent, underserved, and underrepresented communities. AI assurance in this regard requires the sustained, conscientious efforts of relevant designers, developers, and implementors to incorporate human ethics, such as fairness, equity, equality, nondiscrimination, autonomy, and transparency into their designs, services, and products that operate in the infosphere.

AI-facilitated medicine can ameliorate as well as exacerbate healthcare access and delivery disparities. A survey of patient participation in telemedicine during the COVID-19 pandemic, for example, identified racial/ethnic, sex, age, language, and socioeconomic differences in accessing telemedicine for primary care and specialty ambulatory care. If not addressed, such differences may increase existing inequities in care among vulnerable populations (Eberly et al., 2020). Healthcare disparities may occur as broader social features, such as shared distrust of digital devices or of the medical community. Health and healthcare disparities also may reflect patients' inability to pay for healthcare, regardless of the how it is delivered (Ramsetty and Adams, 2020; Valdez et al., 2021). In a 2011–2017 study, for example, the US Veterans Health Administration used administrative data entered via a special portal (My Healthy Vet) to analyze PHRs and to examine demographic characteristics and racial/ethnic differences in portal registration and tool use among veterans with HIV. The data indicated that racial minorities may have been less likely to use PHRs for various reasons, including privacy concerns, lower education levels, and limited internet access (Javier et al., 2019). Analyses of information carried by social media indicate effectiveness in reducing mental health disparities for some marginalized population. For example, as have other groups, some transgender patients have been found to use social media for building community and creating fora for discussions and sharing information. (Grossman et al., 2019; Haimson, 2019).

12.6 Conclusion

The growth of predictive data analytics and the simultaneous growth in the availability of interoperable AI-enabled devices offer opportunities to mitigate healthcare disparities currently endemic in indigent, underrepresented, and underserved communities. However, use of these devices can exacerbate inequities as well as ameliorate them. Thus a variety of policy imperatives can be identified relative to ethical AI. For example, to govern the infosphere, the European Union Commission on AI has recommended seven basic regulatory requirements (Terry, 2019, p. 38): 1) human agency and oversight, 2) technical robustness and oversight, 3) privacy and data governance, 4) transparency, 5) diversity, nondiscrimination and fairness, 6) environmental and social well-being, and 7) accountability.

However, the development of the infosphere in the US is essentially left in the private commercial sector of the economy, and healthcare in the US is a market-based commodity, acquired and traded in a third-party payer system. Government agencies have limited powers to control and manage the acquisition and distribution of healthcare. Nevertheless, national, state, and local governments can employ AI assurance measures and formulate policies and implement programs that address the healthcare needs of specified communities. To make healthcare more available to rural communities, for example, the governments can work together to expand the national electric grid and develop a high-speed internet infrastructure to facilitate the distribution of telemedicine and the dissemination of health information. In addition, US national and state governments can promote access to affordable health insurance coverage available under the Affordable Care Act by increasing the number of health information exchanges and by expanding enrollment periods and eligibility in Medicare and Medicaid programs, by increasing enrollments in programs such as the Children's Health Insurance Program, by increasing tax credits to help reduce healthcare costs among vulnerable populations, and by developing policies and programs to control and lower drug costs. In addition, governments can create participatory multimedia campaigns to curb the spread of misinformation and to overcome, for example, cultural, racial, ethnic, language,

age, and disability barriers to access and participation in healthcare as it affects individual and societal wellbeing.

By themselves, human and AI agents operating in an IoT-sustained infosphere cannot solve the multiple forms of marginalization that affect the health and wellbeing of particular communities. They require actively engaged and empathetic governance to address complex socioeconomic issues and solve complex accessibility and distribution problems. In collaborative environments, where healthcare provision depends on the synergy of human and AI actors, AI assurance and regulatory models must not only address the ethical conduct of medical practitioners, but also the professional conduct of informaticists, software developers, and device vendors.

References

American Medical Association (AMA), 2020. Code of Medical Ethics overview. Available from: https://www.ama-assn.org/delivering-care/ethics/code-medical-ethics-overview.

American Medical Association (AMA), 2021. Artificial Intelligence in Medicine. Available from: Augmented Intelligence (AI) (ama-assn.org).

American Medical Informatics Association (AMIA), 2013. AMIA's code of professional and ethical conduct. Journal of the American Medical Informatics Association 20 (1), 141–143.

Aristotle. Circa 335-323 BCE, 1984. Metaphysics. In: Barnes, J. (Ed.), The Complete Works of Aristotle. Revised Oxford Translation. In: Princeton Bollingen Series LXXI, Part 2, vol. 2. Princeton University Press, pp. 1552–1728.

Aristotle. Circa 340 BCE, 1984. Nicomachean Ethics. In: Barnes, J. (Ed.), The Complete Works of Aristotle. Revised Oxford Translation. In: Princeton Bollingen Series LXXI, Part 2, vol. 2. Princeton University Press, pp. 1729–1867.

Aristotle. Circa 350 BCE, 1984. On the Soul. In: Barnes, J. (Ed.), The Complete Works of Aristotle. Revised Oxford Translation. In: Princeton Bollingen Series LXXI, Part 2, vol. 1. Princeton University Press, Princeton, pp. 641–692.

Batarseh, F., Freeman, L., Huang, C-H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8, 1–30. https://doi.org/10.1186/s40537-021-00445-7.

Boulding, K., 1956. General systems theory—the skeleton of a science. Management Science 2 (3), 197–208.

Boulding, K.E., 1966. The economics of the coming spaceship Earth. In: Jarrett, H. (Ed.), Environmental Quality in a Growing Economy. Resources for the Future/Johns Hopkins University Press, Baltimore, MD, pp. 3–14.

Boulding, K.E., 1970. Economics as a Science. McGraw-Hill, New York.

Boulding, K.E., 1980. Equilibrium, entropy, development, and autopoiesis: towards a disequilibrium Economics. Eastern Economics Journal 6 (3), 179–188.

Bynum, T.W., 1985. Editor's introduction. In: Special Issue: Computers and Ethics. Metaphilosophy 16 (4), 263–265.
Bynum, T.W., 2000. The foundation of computer ethics. Computers & Society 30 (6), 6–13.
Bynum, T.W., 2010. The historical roots of information and computer ethics. In: Florid, L. (Ed.), The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, Cambridge, pp. 20–40.
Cath, C., 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A 376 (20180080), 1–8. https://doi.org/10.1098/rsta.2018.0080.
Coiera, Henrico W.M., 1996. Artificial intelligence in medicine: the challenges ahead. Journal of the American Medical Informatics Association 3 (6), 363–366.
Computer Ethics Institute (CEI), 2011. The Ten Commandments of Computer Ethics. Available from: http://cpsr.org/issues/ethics/cei/.
DeCamp, M., Lindvall, C., 2020. Latent bias and the implementation of artificial intelligence in medicine. Journal of the American Medical Informatics Association 27 (12), 2020–2023.
Department of Health and Human Services (DHHS), 1991. Federal Policy for the Protection of Human Subjects ('Common Rule'). Available from: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.
Department of Health and Human Services (DHHS), 2016. Federal Policy for the Protection of Human Subjects ('Common Rule'). Available from: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.
Eberly, L.A., Kallan, M.J., Julien, H.M., Haynes, N., Khatana, S.A.M., Nathan, A.S., Snider, C., Chokshi, N.P., Eneanya, N.D., Takvorian, S.U., Anastos-Wallen, R., Chaiyachat, K., Ambrose, M., O'Quinn, R., Seigerman, M., Goldberg, L.R., Leri, D., Choi, K., Gitelman, Y., Kolansky, D.M., Cappola, T.P., Ferrari, V.A., Hanson, C.W., Deleener, M.E., Adusumalli, S., 2020. Patient characteristics associated with telemedicine access for primary and specialty ambulatory care during the COVID-19 pandemic. JAMA Network Open 3 (12), e2031640. https://doi.org/10.1001/jamanetworkopen.2020.31640 (Reprinted December 29, 2020). Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2774488.
European Data Protection Supervisor (EDPS), 2014. Mauritius declaration on the Internet of things. In: 36th International Conference of Data Protection and Privacy Commissioners. Available from: Mauritius | European Data Protection Supervisor (europa.eu).
Ferryman, K., 2020. Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. Journal of the American Medical Informatics Association 27 (12), 2016–2019.
Floridi, L., 1999. Information ethics: on the philosophical foundations of computer ethics. Ethics and Computer Technology 1 (1), 37–56.
Floridi, L., 2010a. The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, Cambridge.
Floridi, L., 2010b. Information ethics. In: Floridi, L. (Ed.), The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, Cambridge, pp. 77–100.

Floridi, L., 2018. Soft ethics, the governance of the digital and the General Data Protection Regulation. Philosophical Transactions of the Royal Society A 376, 20180081. https://doi.org/10.1098/rsta.2018.0081.

Floridi, L., Sanders, J.W., 2004. On the morality of artificial agents. Minds and Machines 14 (3), 349–379.

Floridi, L., Sanders, J.W., 2005. Internet ethics: the constructionist values of homo poieticus. In: Cavilier, R. (Ed.), The Impact of the Internet on Our Moral Lives. SUNY Press, Albany, NY, pp. 195–214.

Food and Drug Administration (FDA), 2020. CFR - Code of Federal Regulations Title 21 (revised as of April 1, 2020). Available from: CFR - Code of Federal Regulations Title 21 (fda.gov).

Food and Drug Administration (FDA), 2021a. Artificial Intelligence and Machine Learning in Software as a Medical Device. Retrieved 10-jul-201 from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device; https://www.fda.gov/medical-devices.

Food and Drug Administration (FDA), 2021b. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Available from: AIML_SaMD_Action_Plan (fda.gov).

Gasser, U., Almeida, V.A.F., 2017. A layered model for AI governance. IEEE Internet Computing 21 (6), 58–62.

Gianfrancesco, M.A., Tamang, S., Yazdany, J., Schmajuk, G.G., 2018. Potential biases in machine learning algorithms using electronic health record data. AMA Internal Medicine 178 (11), 1544–1547. Available from: Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data (nih.gov).

Goddard, K., Roudsari, A., Wyatt, J.C., 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19, 121e127. Available from: https://academic.oup.com/jamia/article/19/1/121/732254.

Grossman, L.V., Masterson Creber, R.M., Benda, N.C., Wright, D., Vawdrey, D.K., Ancker, J.S., 2019. Interventions to increase patient portal use in vulnerable populations: a systematic review. Journal of the American Medical Informatics Association 26 (8–9), 855–870.

Haimson, O.L., 2019. Mapping gender transition sentiment patterns via social media data: toward decreasing transgender mental health disparities. Journal of the American Medical Informatics Association 26 (8–9), 749–758.

Henry, T.A., 2021. 7 terms doctors should know about social determinants of health. Available from: Social Determinants of Health | American Medical Association (ama-assn.org).

Hripcsak, G., Albers, David J., 2018. High-fidelity phenotyping: richness and freedom from bias. Journal of the American Medical Informatics Association 25 (3), 289–294.

International Medical Device Regulators Forum (IMDRF), 2020. Strategic Plan 2021–2025. IMDRF/MC/N39FINAL:2020 (Edition 2). Available from: http://www.imdrf.org/.

International Medical Informatics Association (IMIA), 2016. The IMIA Code of Ethics for Health Information Professional. Available from: https://imia-medinfo.org/wp/wp-content/uploads/2015/07/IMIA-Code-of-Ethics-2016.pdf.

Islam, M.M., 2019. Social determinants of health and related inequalities: confusion and implications. Frontiers in Public Health 7 (11), 11. https://doi.org/10.3389/fpubh.2019.00011. National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6376855/.

Javier, S.J., Troszak, L.K., Shimada, S.L., McInnes, D.K., Ohl, M.E., Avoundjian, T., Erhardt, T.A., Midboe, A.M., 2019. Racial and ethnic disparities in use of a personal health record by veterans living with HIV. Journal of the American Medical Informatics Association 26 (8–9), 696–702.

Kannan, V., Wilkinson, K.E., Varghese, M., Lynch-Medick, S., Willett, D.L., Bosler, T.A., Chu, L., Gates, S.I., Holbein, M.E.B., Willett, M.M., Reimold, S.C., Toto, R.D., 2019. Count me in: using a patient portal to minimize implicit bias in clinical research recruitment. Journal of the American Medical Informatics Association 26 (8–9), 703–713.

Kuiler, E.W., McNeely, C.L., 2018. Federal data analytics in the health domain: an ontological approach to data interoperability. In: Batarseh, F.A., Yang, R. (Eds.), Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence. Elsevier, London, pp. 161–176.

Kuiler, E.W., McNeely, C.L., 2020. Knowledge formulation in the health domain: a semiotics-powered approach to data analytics and democratization. In: Batarseh, F.A., Yang, R. (Eds.), Data Democracy at the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering. Elsevier, London, pp. 127–146.

Kulikowski, C.A., 1996. AIM: quo vadis? Journal of the American Medical Informatics Association 3 (6), 432–433.

Lloyd, D., 1985. Frankenstein's children: artificial intelligence and human value. In: Special Issue: Computers and Ethics. Metaphilosophy 16 (4), 307–318.

Magnan, S., 2017. Social Determinants of Health 101 for Health Care: Five Plus Five. National Academy of Medicine. Available from: Social Determinants of Health 101 for Health Care: Five Plus Five - National Academy of Medicine (nam.edu).

McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A., Shaul, R.Z., 2020. Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. Journal of the American Medical Informatics Association 27 (12), 2024–2027.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2019. A Survey on Bias and Fairness in Machine Learning. USC Information Sciences Institute, Los Angelos. Available from: https://arxiv.org/abs/1908.09635.

Moore, J.H., 1985. What is computer ethics? In: Special Issue: Computers and Ethics. Metaphilosophy 16 (4), 266–275.

Moore, J.H., 1999. Just consequentialism and computing. Ethics and Information Technology 1, 65–69.

National Institute of Health (NIH) National Center for Biotechnology Information (NCBI), 2014. Teaching Seven Principles for Public Health Ethics: Towards a Cur-

riculum for a Short Course on Ethics in Public Health Programmes. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196023/.

National Institute of Standards and Technology (NIST), 2019. U.S. LEADERSHIP IN AI: a Plan for Federal Engagement in Developing Technical Standards and Related Tools. Prepared in response to Executive Order 13859. Available from: U.S. LEADERSHIP IN AI: a Plan for Federal Engagement in Developing Technical Standards and Related Tools (nist.gov).

National Institute of Standards and Technology (NIST), 2020. IoT Device Cybersecurity Guidance for the Federal Government: Establishing IoT Device Cybersecurity Requirements. Special Publication 800-213. Available from: https://www.nist.gov/news-events/news/2020/12/nist-releases-draft-guidance-internet-things-device-cybersecurity.

Peirce, C.S., 1958. Collected papers of C.S. Peirce. 8 vols.. In: Hartshorne, C., Wiess, P., Burks (Eds.). Harvard University Press, Cambridge, MA.

Peirce, C.S., 1997. Pragmatism as a principle and method of right thinking. In: Turrisi, P.A. (Ed.), The 1903 Lectures on Pragmatism. SUNY Press, Albany, NY.

Plato. Circa 375 BCE, 1997. Gorgias. In: Cooper, J.M. (Ed.), Plato: The Complete Works. Hacktt Publishing Company, Indianapolis and Cambridge, pp. 791–869.

Plato. Circa 375 BCE, 1997. Republic. In: Cooper, J.M. (Ed.), Plato: The Complete Works. Hacktt Publishing Company, Indianapolis and Cambridge, pp. 971–1223.

Ramesh, A.N., Kambhampati, C., Monson, J.R.T., Drew, P.J., 2004. Artificial intelligence in medicine. Annual Royal College of Surgeons of England 86, 334–338.

Ramsetty, A., Adams, C., 2020. Impact of the digital divide in the age of COVID-19. Journal of the American Medical Informatics Association 27 (7), 1147–1148.

Rigby, M.J., 2019. Ethical dimensions of using artificial intelligence in health care. AMA Journal of Ethics 21 (2), E121–E124. Available from: Ethical Dimensions of Using Artificial Intelligence in Health Care | Journal of Ethics | American Medical Association (ama-assn.org).

Röösli, E., Rice, B., Hernandez-Boussard, T., 2021. Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. Journal of the American Medical Informatics Association 28 (1), 190–192.

Seife, C., 2006. Decoding the Universe: How the New Science of Information Is Explaining Everything from Our Brains to Black Holes. Viking, the Penguin Group, New York.

Solar, O., Irwin, A., 2010. A Conceptual Framework for Action on the Social Determinants of Health. Social Determinants of Health Discussion Paper 2 (Policy and Practice). WHO Document Production Services, Geneva, Switzerland. Available from: WHO Social Determinants of Health Conceptualframeworkforactiononsdh_eng.pdf.

Tavani, H.T., 2010. The foundationalist debate in computer ethics. In: Floridi, L. (Ed.), The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, Cambridge, pp. 251–270.

Terry, N., 2019. Of regulating healthcare AI and robots. Yale Journal of Health Policy, Law, and Ethics 18 (3), 1–51. Available from: https://ssrn.com/abstract=3321379 or http://dx.doi.org/10.2139/ssrn.3321379.

United Nations Educational, Scientific and Cultural Organization (UNESCO), 2019a. Artificial Intelligence: Examples of Ethical Dilemmas. Available from: Artificial Intelligence: examples of ethical dilemmas (unesco.org).

United Nations Educational, Scientific and Cultural Organization (UNESCO), 2019b. Preliminary Study on the Ethics of Artificial Intelligence. Available from: Preliminary study on the Ethics of Artificial Intelligence - UNESCO Digital Library.

United Nations Educational, Scientific and Cultural Organization (UNESCO), 2021. Elaboration of a Recommendation on the Ethics of Artificial Intelligence. Available from: Elaboration of a Recommendation on the ethics of artificial intelligence (unesco.org).

United States Congress, 2020. H.R.1668 - Internet of Things Cybersecurity Improvement Act of 2020. Available from: https://www.congress.gov/bill/116th-congress/house-bill/1668.

Valdez, R.S., Rogers, C.C., Claypool, H., Trieshmann, L., Frye, O., Wellbeloved-Stone, C., Kushalnagar, P., 2021. Ensuring full participation of people with disabilities in an era of telehealth. Journal of the American Medical Informatics Association 28 (2), 389–392.

Veinot, T.C., Ancker, J.S., Bakken, S., 2019. Health informatics and health equity: improving our reach and impact. Journal of the American Medical Informatics Association 26 (8–9), 689–695.

Wiener, N., 1948. Cybernetics: or Control and Communication in the Animal and the Machine. MIT Press, Cambridge, MA.

Wiener, N., 1954. The Human Use of Human Beings, 2nd. rev. ed. Doubleday Anchor, New York.

World Medical Association (WMA), 2018. WMA Declaration Of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

This page intentionally left blank

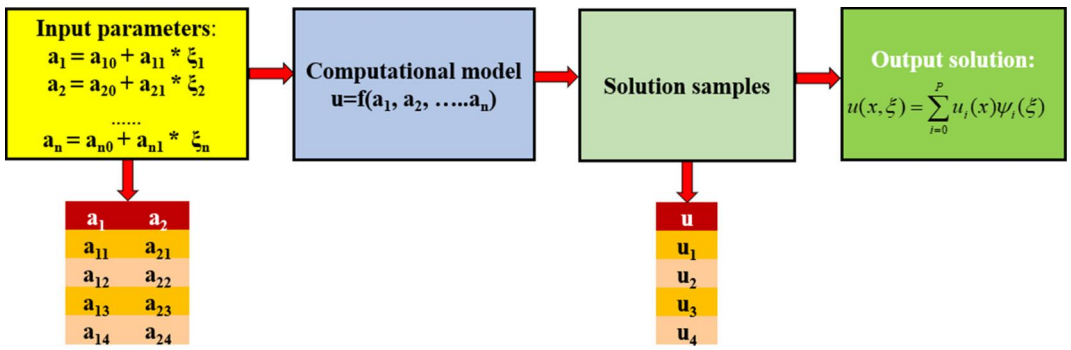
13

Recent advances in uncertainty quantification methods for engineering problems

Dinesh Kumar[a,c], Farid Ahmed[b], Shoaib Usman[c], Ayodeji Alajo[c], and Syed Bahauddin Alam[c]

[a]*Institut National de Recherche en Informatique et en Automatique, Palaiseau, France* [b]*Nuclear Science and Engineering, Military School of Science and Technology, Dhaka, Bangladesh* [c]*Nuclear Engineering and Radiation Science, Missouri University of Science and Technology, Rolla, MO, United States*

Graphical abstract



Abstract

In the last few decades, uncertainty quantification (UQ) methods have been used widely to ensure the robustness of engineering designs. This chapter aims to detail recent advances in popular uncertainty quantification methods used in engineering applications. This chapter describes the two most popular meta-modeling methods for uncertainty quantification suitable for engineering applications: polynomial chaos method and Gaussian process. Furthermore, the UQ methods are applied to an engineering test problem under multiple uncer-

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00027-5

tainties. The test problem considered here is a supersonic nozzle under operational uncertainties. For the deterministic solution, an open-source computational fluid dynamics (CFD) solver SU2 is used. The UQ methods are developed in Matlab® and are further combined with SU2 for the uncertainty and sensitivity estimates. The results are presented in terms of the mean and standard deviation of the output quantities.

Keywords

Uncertainty quantification, polynomial chaos, reliable design, sensitivity analysis, Kriging, Gaussian Process

Highlights

- The chapter updates recent advances in popular uncertainty quantification methods used in engineering applications
- This chapter highlights specific application of supersonic nozzle
- Use of uncertainty quantification in a high dimensional test case is presented
- Efficient approach based on combining polynomial chaos and Kriging is used
- UQ methods are combined with CFD solver SU2 for statistical analysis
- Stochastic results are discussed in terms of mean and standard deviation

13.1 Introduction

In industrial applications, during the operation of engineering devices, several properties and parameters of the components change with time. The material properties of its components vary continuously due to several operational factors. The failure point information of the weak components in an industrial application is very useful. The safety of the engineering device under consideration should always be of utmost concern for the manufacturers while designing the device. Using statistical information, the designer can evaluate the safety margin or make the failure design margin smaller than other components so that the impact of the weak component can be minimized. With advancements in computer hardware and numerical algorithms, computational tools are used to design advanced and high-performance engineering components in almost all engineering fields. For example, aircraft, high-speed car manufacturers, sports, naval ship design-

ers, etc., use computational fluid dynamics (CFD) to simulate fluids over the device. These computational tools are also used for thermal and structural analysis to simulate and detect faults, cracks, and failure of the devices by almost all industries.

Uncertainties are an inherent part of the computing systems concerning real-world applications (Oberkampf and Trucano, 2002; Roy and Oberkampf, 2011; Beyer and Sendhoff, 2007; Schuëller and Jensen, 2008; Wiener, 1938; Xiu and Karniadakis, 2002; Smith, 2013; Kumar et al., 2021a, 2020b; Kabir et al., 2021b,a, 2020; Kumar et al., 2021b, 2020a, 2019). Two physical experiments can never produce the same results, as several of the system parameters are not known properly and have uncertainties. When the same system is modeled using computational tools and mathematical equations, the input and system parameters are provided with constant values to predict the results without dealing with the uncertainties in the input parameters. Almost all aspects of engineering modeling and design are affected by these uncertainties. Engineers and researchers have always encountered issues related to uncertainties in terms of design reliability and robustness. By understanding sources of uncertainties and quantifying them, one can estimate confidence in the system outputs. In mathematical modeling, uncertainties are usually encountered in initial conditions, boundary conditions, material properties, weather conditions, and manufacturing tolerances.

Uncertainty quantification (UQ) is the field of detecting, describing, quantifying, and managing uncertainties in computational designs of real-world systems. In UQ, the system response is estimated in a stochastic way by combining the deterministic solver with statistical tools. UQ methods are statistical tools to assess safety margins in the system responses when computer simulations are used to design an engineering device. UQ methods address the problems associated with incorporating system parameters variability and stochastic behavior into systems analyses. Computer simulations answer what happens when the system under consideration is subjected to a set of input parameters. However, UQ expands this question and answers what will happen when the system is subjected to a range of variability in the input parameters. UQ combines mathematics, statistics, and

engineering. Generally speaking, UQ methods predominantly treat the system to be studied as a closed system (like a black box), and an extensive understanding of the system's inner functioning is not required. The UQ methods only need information about the input parameters of the model and model responses to estimate probabilistic model responses. In the recent past, uncertainty quantification and management are considered as significant elements in risk management (the system can fail or be damaged if it does not meet the design targets) of industrial designs (Oberkampf and Trucano, 2002; Hirsch et al., 2018).

Due to the non-intrusive nature of UQ methods, these methods can be adopted easily by researchers and industries from a wide range of engineering, industrial and financial sectors to achieve the following:

- Understand the uncertainties inherent in the system.
- Predict system responses concerning uncertain inputs.
- Quantify confidence in the system responses.
- Find optimal responses concerning a wide range of inputs.
- Reduce unexpected system failures.
- Implement probabilistic modeling and design processes.
- Predict parametric sensitivity on the model responses.

With increasing computational power and simulation techniques, it became possible to make accurate predictions of real-world systems. Now the challenges in engineering designs are moved toward predicting system behaviors with respect to uncertainties efficiently. Traditional UQ methods based on Monte Carlo (Hammersley, 2013; Rubinstein and Kroese, 2016) usually need a large number of system evaluations. So these methods are restricted to simplified test cases and for research purposes only. Monte Carlo methods are sampling-based methods, and the convergence rate is very slow. In general, large samples (at least 10^4) are needed to predict statistical quantities accurately. Alternatively, the literature proposes several sampling schemes, such as Latin hypercube, sparse sampling, clustered sampling, and stratified sampling to accelerate the convergence. However, Monte Carlo methods have not gained massive popularity due to their expensive computational cost. For large-scale problems and real-world engineering applications, more recent methods based on machine

learning approaches, such as polynomial chaos method (PCM) (Xiu and Karniadakis, 2002; Najm, 2009; Hosder et al., 2007; Ghanem et al., 2017), Gaussian process (or Kriging) (Quinonero-Candela and Rasmussen, 2005; Bastos and O'Hagan, 2009), support vector machine (SVM) (Awad and Khanna, 2015; Smola and Schölkopf, 2004), polynomial-chaos-Kriging (or PC-Kriging) (Schobi et al., 2015; Wang et al., 2019) are proposed in literature and are applied to several diverse applications. Polynomial chaos and Gaussian process models are seen as leading approaches for stochastic and robustness analyses of very complex engineering applications. PC-Kriging is a result of combining polynomial chaos and Kriging methods. These approaches are discussed in detail and are applied to an engineering application in the sections that follow.

To understand the potential influence of input parameters on system outputs, the sensitivity analysis (SA) method is used (Saltelli, 2002; Sudret, 2008). Various methods for global sensitivity analysis, such as linear regression and variance-based analysis are discussed for sensitivity estimation. A standard method of estimating system responses is using Sobol' indices-based global sensitivity analyses. Various meta-modeling methods can compute Sobol indexes, including Monte Carlo, graphical models, Kriging, and support vector machine approaches. In recent years, surrogate models that calculate Sobol' indices have gained considerable attention. The first step in this approach is to construct a surrogate model using the design of experiments (DOE). Furthermore, this surrogate model is used to estimate many model responses to compute Sobol' indices. Computing model responses from surrogate models are also called data-driven approaches, as several combinations of input parameters are used to explore a wide range of input domains.

13.2 Polynomial chaos method for UQ

In 1938, Wiener proposed the polynomial chaos method for dealing with Gaussian distributed uncertainties (Wiener, 1938). Xiu and Karniadakis (Xiu and Karniadakis, 2002) demonstrated the ability to use it with any probability distribution in a detailed analysis. In the last few years, the generalized method has been used in a variety of engineering applications,

including computational fluid dynamics, heat transfer, nuclear reactor design, and structural analysis (Kumar et al., 2020c). Because adding uncertainties increases the computation required to quantify them, early applications dealt with a limited number of uncertainties. With increasing uncertainties, the number of simulations required to quantify the uncertainty grows exponentially using the polynomial chaos method. This is referred to as the dimensionality curse. Numerous improvements have been proposed in literature (Blatman and Sudret, 2011; Liu et al., 2020; Hosder et al., 2007; Kumar et al., 2016; Aremu et al., 2020; Liu and Bellet, 2019) to cope with the curse of dimension and move the UQ process forward. Several researchers also proposed model reduction algorithms (based on principal component analysis) to accelerate the polynomial chaos method.

Numerous applications have used model reduction approaches. However, these approaches mainly were two-step processes and usually were applied in semi-intrusive ways. Thus they were not very straightforward to use for engineering applications where the models can be used as a black-box. Several researchers proposed the idea of sparse sampling. Using sparse sampling schemes (such as Fejer, Clenshaw–Curtis, Conrod–Patterson), the number of simulations can be reduced to achieve the same accuracy as classical polynomial chaos. Blatman and Sudret proposed a theory of sparse polynomial chaos, based on least angle regression in their paper (Blatman and Sudret, 2011; Bourinet, 2018). Based on its principle, this method used a maximum number of polynomial order approximations for a given number of samples and a sparse polynomial chaos expansion (PCE) for a given system response (Kumar et al., 2020c, 2021a,b, 2020b, 2016). Several other researchers also proposed the more or less similar idea of sparse polynomial chaos using different error minimizing schemes. Recently, numerous applications have seen the sparse polynomial chaos approach due to its straightforward usage and faster convergence capability. In this section, some fundamental concepts for the polynomial chaos approach are described (Ghanem et al., 2017). It is relevant to note that the lead author developed this method, and the descriptions have been reported in different studies (Kumar et al., 2020c, 2021a,b, 2020b, 2016) for a range of engineering applications.

Based on a set of orthogonal polynomial basis functions, we can write a stochastic model response for a system under uncertainty as follows:

$$Y = M(\boldsymbol{\xi}) = \sum_{\boldsymbol{b} \in \mathbb{N}^n} a_{\boldsymbol{b}} \psi_{\boldsymbol{b}}(\boldsymbol{\xi}), \tag{13.1}$$

where $\psi_{\boldsymbol{b}}$ is an orthogonal polynomial for multidimensional dimensions, $\boldsymbol{b} = b_1 \dots b_n$ represents an index, and the terms $a_{\boldsymbol{b}}$ are called polynomial coefficients. A PCE is given by Eq. (13.1). A system of n-dimensional input uncertainties is represented by $\boldsymbol{\xi}$ in the above equation. From a set of orthogonal one-dimensional polynomials, we construct the multi-dimensional polynomials $\psi_{\boldsymbol{b}}$ as follows (Kumar et al., 2016):

$$\psi_{\boldsymbol{b}}(\boldsymbol{\xi}) = \psi_{b_1 \dots b_n}(\boldsymbol{\xi}) = \prod_{i=1}^{n} \psi_{b_i}(\xi_i), \tag{13.2}$$

where b_i is the order of the polynomial expansion for the random variable ξ_i.

Extended polynomial expansions usually truncate to a finite number of terms, because higher-order terms are not significant in the system response after a few terms. We truncate PCE into the following to achieve the following degree $|\boldsymbol{b}| = \sum_{i=1}^{n} b_i$ within a given order p (Du, 2019):

$$Y \simeq M_p(\boldsymbol{\xi}) = \sum_{\boldsymbol{b} \in A^{p,n}} a_{\boldsymbol{b}} \psi_{\boldsymbol{b}}(\boldsymbol{\xi}), \; A^{p,n} = \{\boldsymbol{b} \in \mathbb{N}^n : |\boldsymbol{b}| \leq p\}. \tag{13.3}$$

The total number of terms, P (basis functions), equals $\frac{(n+p)!}{n!p!}$ when the number of input uncertainties is n and the highest order of polynomial in PCE is p. Polynomial coefficients can be calculated based on the PCE order and solution samples (system responses using a deterministic solver as black box). One can compute and construct the PCE of a stochastic output. In the PCE, the first term (the zeroth-order term) represents the stochastic response's mean. In addition, one can also compute higher-order statistical moments numerically by using these polynomial coefficients. Computing polynomial coefficients can be done using numerical methods, such as collocation and regression (Kumar et al., 2016).

Once we calculate polynomial coefficients, the mean $E(Y)$ and variance $V(Y)$ of the system output Y can be computed easily as below:

$$E(Y) = a_0; V(Y) = \sum_{i=1}^{P} a_i^2 \psi_i^2, \tag{13.4}$$

where the coefficients a_i and ψ_i are the same as they were defined earlier.

13.3 Gaussian Process or Kriging for UQ

Kriging, also known as Gaussian process modeling, is a statistical method for approximating various functions and computer experiments using Gaussian processes. Kriging is also used as a surrogate model to establish a link between the inputs and outputs of expensive computational models (Zhang and Apley, 2016). The Gaussian process has been used for several machine learning applications related to regression and classification in the last few decades. Krige first developed Kriging method for geostatistical applications in 1951. In addition, it was used in metamodeling and data-driven modeling for numerous applications with noisy data. The model is known as Kriging (after Krige in geostatistics). Using Gaussian processes, Tarantola and Valette designed a Bayesian formulation for inverse problems in geophysics (Tarantola et al., 1982). Based on the work of Williams, Neal, and Rasmussen, the model was proposed to solve regression problems in statistics (O'Hagan, 1978; Hinton et al., 1995; Williams and Rasmussen, 1996) and gained popularity. (Hinton et al., 1995; Williams and Rasmussen, 1996; Gibbs and MacKay, 1997) provides the Bayesian interpretation and detailed description of the model. Machine learning was introduced to Gaussian process in the nineties. As a result of a detailed comparison by Rasmussen (Hinton et al., 1995; Williams and Rasmussen, 1996) of the GP with the most widely used models, the GP started becoming very popular. He showed that GP approaches outperformed other approaches in the vast majority of cases. Using the maximum-likelihood estimation method (MLE), the GP model's learning process involves tuning the covariance parameters to the data. To obtain the prediction and the degree of uncertainty associated with it, given a new input and con-

ditioned on previous observations, one can easily calculate the mean and variance of the predictive distribution. By using the definition of conditional probabilities, we can easily obtain Gaussian distribution based on the GP assumption (Girard, 2004).

The output response of the model M, according to Kriging, is the realization of a Gaussian process. Kriging metamodel $M^K(x)$ of the true model $M(x)$ can be described as

$$M^K(x) = \beta^T f(x) + \sigma^2 Z(x, \omega), \tag{13.5}$$

where $\beta^T f(x)$ is the mean of the Gaussian process, σ^2 is the variance of the process, and $Z(x, \omega)$ is a stationary Gaussian process with a zero mean and unit variance (Du, 2019). The underlying probability space (ω) is defined in terms of a correlation function $R(x_1, x_2; \theta)$ that describes the correlation between two sample points in the output space x_1 and x_2, as well as the hyperparameters θ.

If $y = y_1, y_2, y_3, ..., y_N$ are the outputs of the true model $M(x)$ at sampling points $x = x_1, x_2, x_3, ..., x_N$, the model prediction $M^K(x)$ at a new point x can be estimated using Kriging metamodeling. The gaussian process metamodeling prediction is based on the fact that the prediction y' at the new point x and the responses from the true model y make a joint Gaussian distribution as:

$$\begin{Bmatrix} y' \\ y \end{Bmatrix} = \mathbb{N}_{N+1} \left(\begin{Bmatrix} f^T(x)\beta \\ F\beta \end{Bmatrix}, \sigma^2 \begin{Bmatrix} r(x)^T r(x) \\ R \end{Bmatrix} \right). \tag{13.6}$$

In the above equation, F is the observation matrix with entries $f_j(x_i)$ for $i = 1, 2, 3, ..., N$ and $j = 1, 2, 3, ..., P$, where $f_j(x_i)$ are arbitrary functions at observation points x_i, and β are regression coefficients. The vector $r(x)$ is the cross correlations between the new point x and the known points x_i as

$$\begin{Bmatrix} r_1 \\ r_2 \\ r_3 \\ . \\ . \\ r_N \end{Bmatrix} = \begin{Bmatrix} R(x, x_1, \theta) \\ R(x, x_2, \theta) \\ R(x, x_3, \theta) \\ . \\ . \\ R(x, x_N, \theta) \end{Bmatrix}, \tag{13.7}$$

where R is the correlation matrix at the known points x_i and can be written as

$$R_{ij} = R(x_i, x_j; \theta), \tag{13.8}$$

where $i, j = 1, 2, 3, ..., N$, or

$$R = \begin{Bmatrix} R(x_1, x_1, \theta) & R(x_1, x_2, \theta) & . & . & . & R(x_1, x_N, \theta) \\ R(x_2, x_1, \theta) & R(x_2, x_2, \theta) & . & . & . & R(x_2, x_N, \theta) \\ R(x_3, x_1, \theta) & R(x_3, x_2, \theta) & . & . & . & R(x_3, x_N, \theta) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ R(x_N, x_1, \theta) & R(x_N, x_2, \theta) & . & . & . & R(x_N, x_N, \theta) \end{Bmatrix}. \tag{13.9}$$

Using the conditional distribution properties of the multivariate normal, the mean and the variance of the predictor can be written as

$$E\{y'|y\} = f^T\beta + r^T R^{-1}(y - F\beta), \tag{13.10}$$

$$\sigma_y^2 = \sigma^2(1 - r^T R^{-1} r + u^T (F^T R^{-1} F)^{-1} u), \tag{13.11}$$

where the regression coefficients β and the term u are defined as

$$\beta = (F^T R^{-1} F)^{-1} F^T R^{-1} y, \tag{13.12}$$

$$u = F^T R^{-1} r - f. \tag{13.13}$$

13.4 Polynomial chaos Kriging for UQ

Kriging interpolates local variations in the system response Y as a function of the neighboring design points, whereas PCE closely approximates the regional behavior of Y (Amini et al., 2021). It is possible to obtain more accurate PC-Kriging metamodels by combining local and global approximation techniques. In PC-Kriging, there is an array of orthonormal polynomials that represent the trend and which are defined as follows:

$$M^{PCK}(x) = \sum_{\boldsymbol{b} \in A^{p,n}} a_{\boldsymbol{b}} \psi_{\boldsymbol{b}}(\boldsymbol{\xi}) + \sigma^2 Z(x, \omega), \tag{13.14}$$

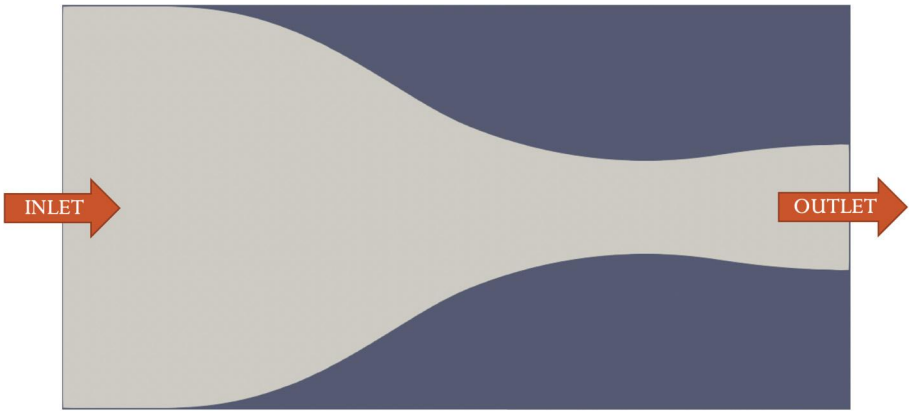


FIGURE 13.1 CD nozzle.

where ψ_b are multivariate orthogonal polynomials concerning the input distributions, and a_b are the corresponding coefficients.

13.5 Uncertainty quantification of a supersonic nozzle

The UQ is crucial for assuring the results produced by mathematical modeling in engineering applications. The standard deviation or variance can be considered a safety bound or a confidence interval around the mean values. Hence, we apply the methods discussed in the previous sections to an engineering application in this section. We analyze the non-ideal supersonic compressible flow within a 2D converging-diverging nozzle shown in Fig. 13.1 (Guardone, 2021). The nozzle is 0.123 m long, with a throat height of 0.0084 m and an inlet height of 0.036 m (Guardone, 2021). This is a test case provided in SU2 for deterministic CFD simulations (Guardone, 2021). It has a simple geometry, where the flow accelerates from subsonic to supersonic speeds. It can be used to investigate compressible flows, where simple ideal gas laws are not enough to describe thermodynamic behavior properly. To determine the performance of the nozzle over a wide range of inlet conditions, CFD simulations are first used to confirm its performance. Then further CFD simulations are combined with uncertainty quantification methods. Additionally, the uncertainty bounds for the nozzle performance in terms of pressure and Mach number, flow density, and

Table 13.1 Flow conditions for C-D Nozzle CFD simulation.

Parameters	Values
Working fluid	Octamethyltrisiloxane
Inlet pressure	904,388 Pa
Inlet temperature	542.13 K
Turbulence model	SST
Gamma value	1.01767
Gas constant	35.17
Critical temperature	565.3609
Critical pressure	1,437,500
μ	1.21409E-05
K_T	0.030542828

temperature fields along the centerline are evaluated with respect to input uncertainties. They are shown with mean and standard deviation values.

13.5.1 Test case description

Octamethyltrisiloxane (MDM), a pressure-sensitive fluid, is used as a working fluid for analyzing non-ideal supersonic compressible flow inside a converging-diverging (CD) nozzle. Table 13.1 presents details about the properties of fluid and flow conditions. This configuration results in a total exhaust pressure ratio of 3.125, which results in a supersonic outflow at Mach number 1.5 (Guardone, 2021). The static pressure applied to the test case's outlet is 200,000 Pa. The computational domain and mesh are depicted in Fig. 13.2. The mesh is composed of 3540 quadrilateral elements and 3660 nodes (Guardone, 2021). At the inlet and outlet boundaries, Riemann boundary conditions based on characteristics are used. Symmetry boundary conditions define symmetry boundaries. By mirroring the flow around the x-axis, the mesh size is reduced, along with the computational cost. On the boundary of a wall, Navier–Stokes adiabatic wall conditions are applied.

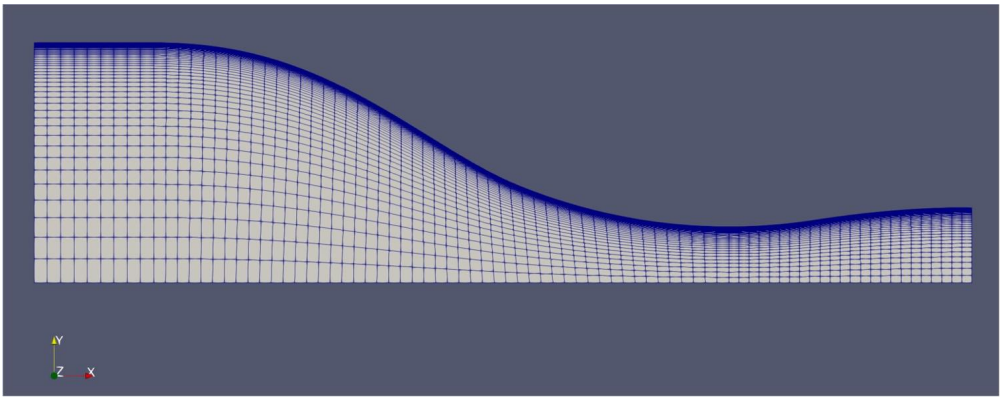


FIGURE 13.2 Computational domain and mesh.

13.5.2 Deterministic results

For deterministic simulations, the SU2 solver is run at the fixed boundary conditions, flow conditions, and fluid properties as described earlier. The total number of iteration was given 1000 so that all solutions and residuals are converged nicely. We post-process the data with Paraview (a multi-platform, open-source data analysis and visualization application). See Fig. 13.3. In Fig. 13.4, solution fields for pressure, temperature, Mach number, and flow density are exhibited for the whole computational domain. Furthermore, these quantities are also shown at the centerline of the nozzle for better understanding and analysis. At the inlet of the nozzle, pressure, temperature, and density of the fluid are at their maximum, and then at their minimum near the outlet. In an inlet, the Mach number can be viewed as a minimum, and at the exit, the Mach number reaches a maximum value.

13.5.3 Description of uncertainties

For the uncertainty analysis, seven input parameters; two from boundary values (inlet temperature and inlet pressure), three from gas properties (specific heat ratio γ, gas constant R and acentric factor ω), and two from the viscosity model (molecular viscosity μ and molecular thermal conductivity K_T) are considered as uncertain. All parameters are considered uniformly distributed. The inlet pressure and acentric factor are assumed to

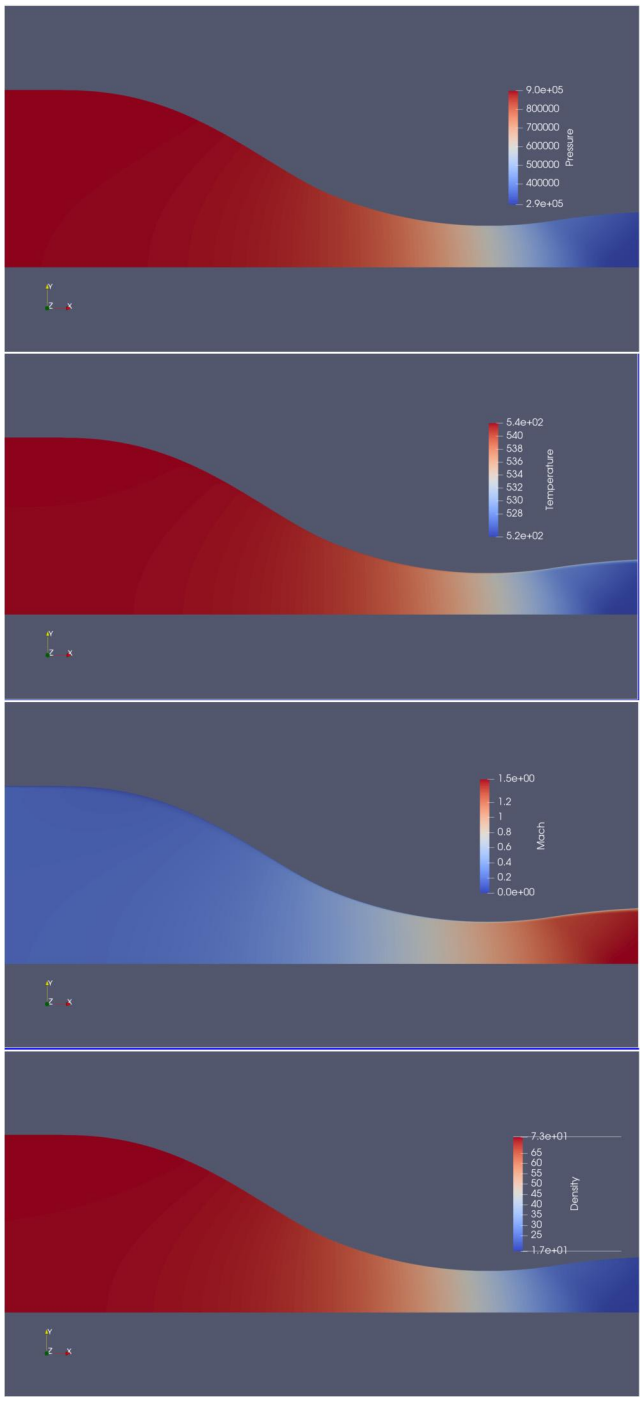


FIGURE 13.3 Numerical results (provided serially in vertical format): pressure, temperature, Mach, and density fields.

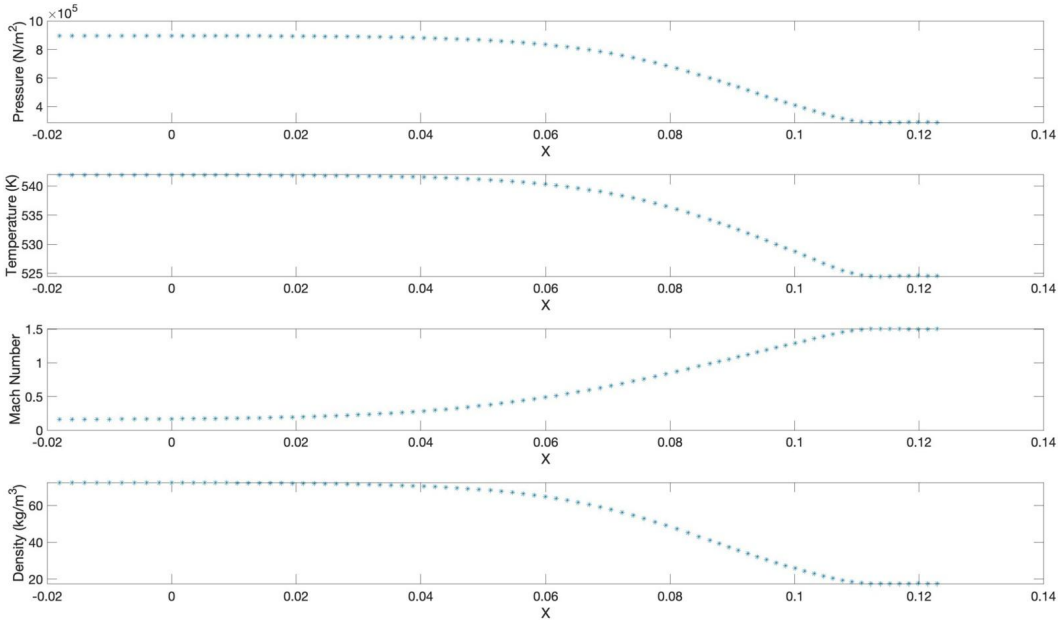


FIGURE 13.4 Computational results: pressure, temperature, Mach number and fluid density at the centerline of the nozzle).

Table 13.2 Input uncertainties for CFD simulations.

Parameters	Values	Uncertainties (%)	Minimum	Maximum
Inlet pressure (Pa)	904,388	5	859,168	949,607
Inlet temperature (K)	542.13	1	536.71	547.55
Gamma value	1.01767	1	1.00749	1.02785
Gas constant	35.17	2	34.47	35.87
μ	1.21409E-05	2	1.18981	1.23837
K_T	0.030542828	2	0.029931971	0.031153684
Acentric factor (ω)	0.524	5	0.498	0.550

have 5% variability from the mean value. Gas constant, molecular viscosity, and molecular thermal conductivity are assumed to vary 2% from their mean values. Minor uncertainties of 1% from their mean values are given to the inlet temperature and Gamma values. All the mean values for these parameters, their uncertainties, and their ranges of variability are described in Table 13.2.

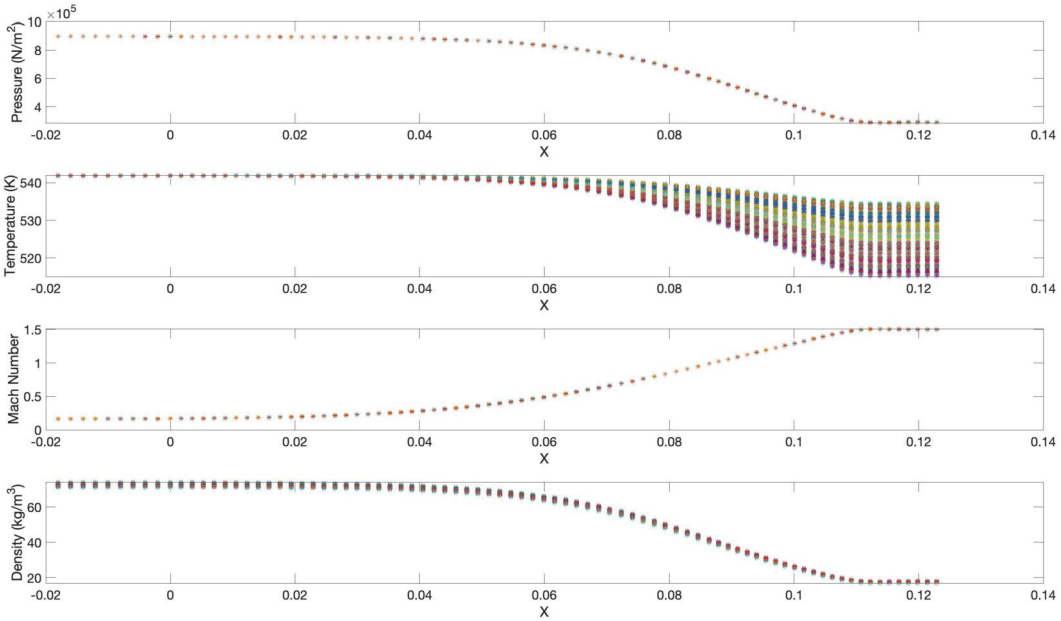


FIGURE 13.5 Solution samples for pressure, temperature, Mach number and fluid density at the centerline of the nozzle).

13.5.4 Uncertainty analysis

As described in the previous section, the PC-Kriging method is used here to estimate the combined impact of all input uncertainties on the system responses of the CD nozzle. Usually, in the regression-based polynomial chaos method, a total of 240 CFD samples (for PC order 3 and 7 input uncertainties) will be required to estimate the statistical quantities (mean and standard deviance) of the output accurately (see (Kumar et al., 2016)). Here to construct the PC-Kriging-based surrogate model, only 100 CFD simulations are used. For input parameters, 100 designs of experiments are constructed using the Sobol sequence-based sampling technique. In Fig. 13.5, the CFD solutions for pressure, temperature, Mach number, and fluid density along the nozzle centerline are shown for all 100 samples. It can be seen that pressure and Mach number are not varying much with the input uncertainties. However, minor variations can be seen in density with the input variations. The most significant variations can be seen for the temperature field. In Fig. 13.6, the mean and standard deviation are shown for

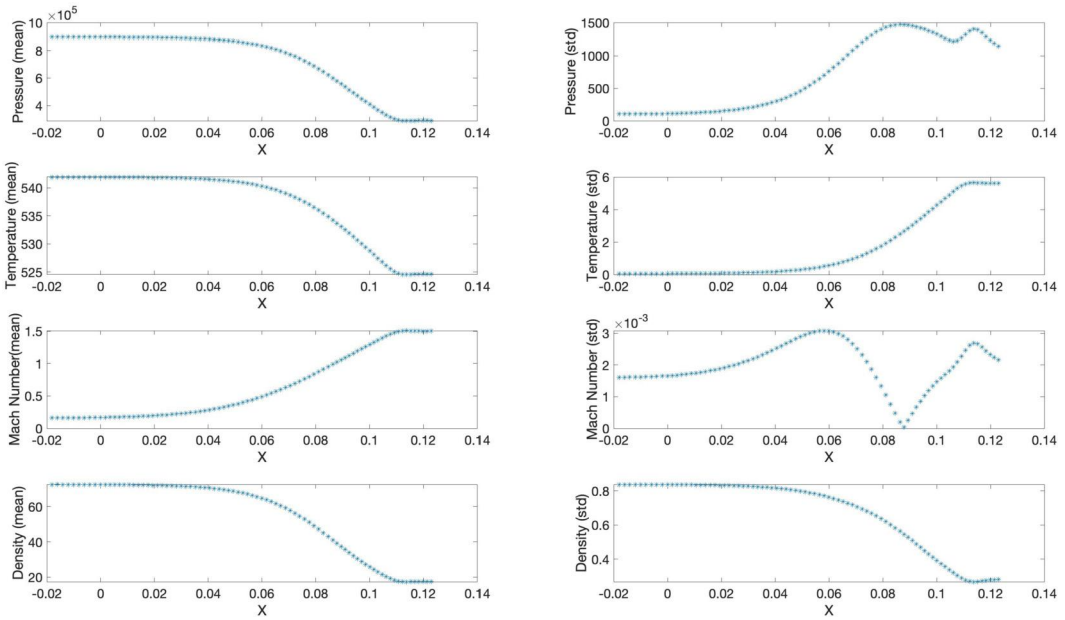


FIGURE 13.6 Mean and standard deviation for pressure, temperature, Mach number and fluid density at the centerline of the nozzle).

all these quantities. These values are calculated from the PC-Kriging-based surrogate model. The mean values behavior is similar to the deterministic solutions. For pressure, temperature, and Mach number, the standard deviation values are higher at the outlet. That means the highest fluctuations are at the nozzle outlet. However, the standard deviation for fluid density is seen lower at the nozzle outlet.

It is also important to highlight that this developed uncertainty method can be applied to other domains such as nuclear engineering in terms of safety assessment of advanced reactor system (Kumar et al., 2021a). In addition, the authors also utilized this methodology to understand and evaluate the uncertainties in composite materials (Kumar et al., 2021b).

13.6 Conclusions

In this work, first, we describe the two most popular meta-modeling methods (Polynomial Chaos and Kriging methods) suitable for uncertainty

quantification in engineering applications. Furthermore, to increase the efficiency, the polynomial chaos and Kriging methods are combined and used for an engineering test problem under multiple uncertainties. A 2D supersonic converging-diverging nozzle is considered for the analysis where the multi-physics CFD solver SU2 is used for deterministic solutions. The UQ methods (polynomial chaos, Kriging, and PC-Kriging) are developed in Matlab® and are further combined with SU2 for uncertainty quantification. The standard deviation can be considered as a safety bound or a confidence interval around the mean values. Hence, for assurance in making crucial decisions, the results are discussed in terms of the mean and standard deviation of the output quantities, i.e., pressure, temperature, Mach number, and fluid density.

Future work will focus on its application in multiscale modeling of composite accident-tolerant nuclear fuels with Sic/Sic claddings for small modular reactor (SMR) applications.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. OAC-1919789.

References

Amini, A., Abdollahi, A., Hariri-Ardebili, M., Lall, U., 2021. Copula-based reliability and sensitivity analysis of aging dams: adaptive Kriging and polynomial chaos Kriging methods. Applied Soft Computing 107524.

Aremu, O.O., Hyland-Wood, D., McAree, P.R., 2020. A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data. Reliability Engineering & Systems Safety 195, 106706.

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines. Springer, pp. 67–80.

Bastos, L.S., O'Hagan, A., 2009. Diagnostics for Gaussian process emulators. Technometrics 51, 425–438.

Beyer, H.G., Sendhoff, B., 2007. Robust optimization – a comprehensive survey. Computer Methods in Applied Mechanics and Engineering 196, 3190–3218.

Blatman, G., Sudret, B., 2011. Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of Computational Physics 230, 2345–2367.

Bourinet, J.M., 2018. Reliability analysis and optimal design under uncertainty-Focus on adaptive surrogate-based approaches. PhD thesis. Université Clermont Auvergne.

Du, X., 2019. Efficient uncertainty propagation for model-assisted probability of detection and sensitivity analysis via metamodeling and multifidelity methods. Ph.D. thesis. Iowa State University.
Ghanem, R., Higdon, D., Owhadi, H., 2017. Handbook of Uncertainty Quantification, vol. 6. Springer.
Gibbs, M., MacKay, D.J., 1997. Efficient implementation of Gaussian processes.
Girard, A., 2004. Approximate Methods for Propagation of Uncertainty with Gaussian Process Models. University of Glasgow, United Kingdom.
Guardone, A., 2021. Non-ideal compressible flow in a supersonic nozzle. https://su2code.github.io/tutorials/NICFD_nozzle/.
Hammersley, J., 2013. Monte Carlo Methods. Springer Science & Business Media.
Hinton, G.E., Dayan, P., Frey, B.J., Neal, R.M., 1995. The "wake-sleep" algorithm for unsupervised neural networks. Science 268, 1158–1161.
Hirsch, C., Wunsch, D., Szumbarksi, J., Laniewski-Wollk, L., Pons-Prats, J., 2018. Uncertainty management for robust industrial design in aeronautics. Notes on Numerical Fluid Mechanics and Multidisciplinary Design 140.
Hosder, S., Walters, R., Balch, M., 2007. Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. In: 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, p. 1939.
Kabir, H.D., Khosravi, A., Kavousi-Fard, A., Nahavandi, S., Srinivasan, D., 2021a. Optimal uncertainty-guided neural network training. Applied Soft Computing 99, 106878.
Kabir, H.D., Khosravi, A., Mondal, S.K., Rahman, M., Nahavandi, S., Buyya, R., 2021b. Uncertainty-aware decisions in cloud computing: foundations and future directions. ACM Computing Surveys (CSUR) 54, 1–30.
Kabir, H.D., Khosravi, A., Nahavandi, D., Nahavandi, S., 2020. Uncertainty quantification neural network from similarity and sensitivity. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.
Kumar, D., Alam, S., Ridwan, T., Goodwin, C.S., 2021a. Quantitative risk assessment of a high power density small modular reactor (SMR) core using uncertainty and sensitivity analyses. Energy 227, 120400.
Kumar, D., Alam, S., Sjöstrand, H., Palau, J., De Saint Jean, C., 2019. Influence of nuclear data parameters on integral experiment assimilation using Cook's distance. In: EPJ Web of Conferences. EDP Sciences, p. 07001.
Kumar, D., Alam, S., Sjöstrand, H., Palau, J., De Saint Jean, C., 2020a. Nuclear data adjustment using Bayesian inference, diagnostics for model fit and influence of model parameters. In: EPJ Web of Conferences. EDP Sciences, p. 13003.
Kumar, D., Alam, S., Vučinić, D., Lacor, C., 2020b. Uncertainty quantification and robust optimization in engineering. In: Advances in Visualization and Optimization Techniques for Multidisciplinary Research. Springer, pp. 63–93.
Kumar, D., Koutsawa, Y., Rauchs, G., Marchi, M., Kavka, C., Belouettar, S., 2020c. Efficient uncertainty quantification and management in the early stage design of composite applications. Composite Structures, 112538.
Kumar, D., Marchi, M., Alam, S.B., Kavka, C., Koutsawa, Y., Rauchs, G., Belouettar, S., 2021b. Multi-criteria decision making under uncertainties in composite materials selection and design. Composite Structures, 114680.

Kumar, D., Raisee, M., Lacor, C., 2016. An efficient non-intrusive reduced basis model for high dimensional stochastic problems in CFD. Computers & Fluids 138, 67–82.
Liu, H., Jiang, C., Xiao, Z., 2020. Efficient uncertainty propagation for parameterized p-box using sparse-decomposition-based polynomial chaos expansion. Mechanical Systems and Signal Processing 138, 106589.
Liu, K., Bellet, A., 2019. Escaping the curse of dimensionality in similarity learning: efficient Frank–Wolfe algorithm and generalization bounds. Neurocomputing 333, 185–199.
Najm, H.N., 2009. Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. Annual Review of Fluid Mechanics 41, 35–52.
Oberkampf, W.L., Trucano, T.G., 2002. Verification and validation in computational fluid dynamics. Progress in Aerospace Sciences 38, 209–272.
O'Hagan, A., 1978. Curve fitting and optimal design for prediction. Journal of the Royal Statistical Society, Series B, Methodological 40, 1–24.
Quinonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research 6, 1939–1959.
Roy, C.J., Oberkampf, W.L., 2011. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Computer Methods in Applied Mechanics and Engineering 200, 2131–2144.
Rubinstein, R.Y., Kroese, D.P., 2016. Simulation and the Monte Carlo Method, vol. 10. John Wiley & Sons.
Saltelli, A., 2002. Sensitivity analysis for importance assessment. Risk Analysis 22, 579–590.
Schobi, R., Sudret, B., Wiart, J., 2015. Polynomial-chaos-based Kriging. International Journal for Uncertainty Quantification 5.
Schuëller, G., Jensen, H.A., 2008. Computational methods in optimization considering uncertainties – an overview. Computational Methods in Applied Mechanical Engineering 198, 2–13.
Smith, R.C., 2013. Uncertainty Quantification: Theory, Implementation, and Applications, vol. 12. Siam.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222.
Sudret, B., 2008. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering & Systems Safety 93, 964–979.
Tarantola, A., Valette, B., et al., 1982. Inverse problems= quest for information. Journal of Geophysics 50, 159–170.
Wang, F., Xiong, F., Chen, S., Song, J., 2019. Multi-fidelity uncertainty propagation using polynomial chaos and Gaussian process modeling. Structural and Multidisciplinary Optimization 60, 1583–1604.
Wiener, N., 1938. The homogeneous chaos. American Journal of Mathematics 60, 897–936.
Williams, C.K., Rasmussen, C.E., 1996. Gaussian processes for regression.
Xiu, D., Karniadakis, G.E., 2002. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM Journal on Scientific Computing 24, 619–644.
Zhang, N., Apley, D.W., 2016. Brownian integrated covariance functions for Gaussian process modeling: sigmoidal versus localized basis functions. Journal of the American Statistical Association 111, 1182–1195.

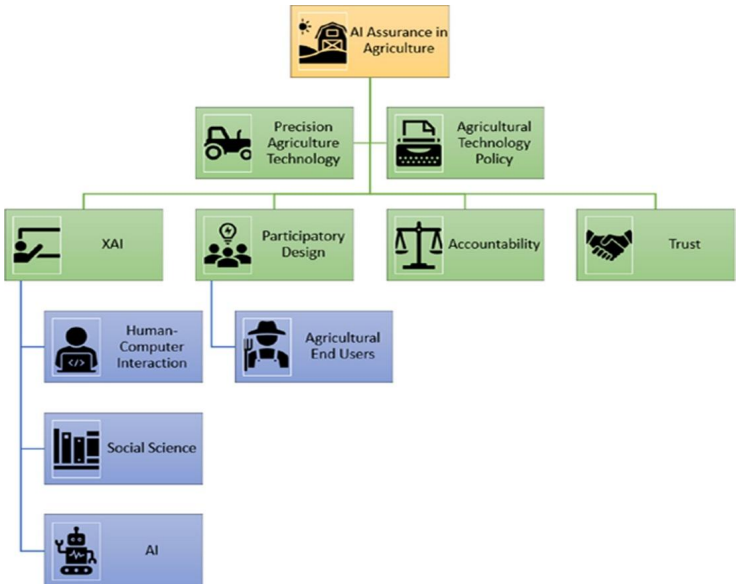
14

Socially responsible AI assurance in precision agriculture for farmers and policymakers

Brianna B. Posadas[a], Ayorinde Ogunyiola[b], and Kim Niewolny[a]

[a]*Virginia Polytechnic Institute and State University, Blacksburg, VA, United States*
[b]*Purdue University, West Lafayette, IN, United States*

Graphical abstract



Abstract

As one solution to feeding a growing population with finite resources, some farmers, researchers, and agricultural technology providers (ATPs) have turned

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00028-7

to precision agriculture (PA). PA is the practice of mapping out precise input application to maximize the yield. To do this, ATPs collect input and output data from farmers and use Artificial Intelligence and machine learning to build prescription maps, which farmers can program farm equipment to follow. The use of PA has allowed farmers to use less resources, which saves money and reduces environmental impact. However, technology is a two-sided coin, benefiting both end-users, the farmers, and ATPs differently. In agriculture, power asymmetry has been cited as a critical issue existing between farmers and ATPs,and this impacts farmers negatively. For farmers to deploy and have more control over data decision-making on their farms, AI assurance methods need to be integrated into their technologies. There are currently a few studies on this subject in agriculture, but many do not involve agricultural end-users or fall short of meeting the needs of the end-users. If end-users and policymakers are not able to understand how their data is collected and used in the agricultural AI models, they will not be able to make educated decisions about their work. This chapter proposes solutions to benefit all agricultural end-users, including prompting the use of participatory design and adopting more user-centered principles when integrating AI assurance models into agricultural technologies.

Keywords

Precision agriculture, explainable AI, public policy, user-centered design

Highlights

- Currently, there are less than 10 studies of AI assurance in agriculture. Of these studies, few test their systems with agricultural experts; none test their systems with agricultural workers
- This is significant, because as agriculture becomes more automated and technologically advanced, the end users who need to understand the "black box" of AI system are overwhelmingly not ML experts
- Having this "black box" makes it difficult for agricultural workers to trust AI systems and makes it difficult for policymakers to protect the interests of agricultural stakeholders
- This chapter proposes several recommendations for more accessible XAI agricultural systems, including utilizing participatory design, designing for different end users, and having programmers be transparent and upfront about data use and privacy

14.1 Introduction

As one solution to feeding a growing population with finite resources, some farmers, researchers, and agricultural technology providers (ATPs) have turned to Precision Agriculture (PA) (Chaterji et al., 2020; Ryan, 2019). PA is the practice of mapping out precise input application to maximize the yield. Industrial agricultural fields can range from 50 hectares (ha) to over 1000 ha in size, which means the characteristics of the soil can vary widely from one side to the other, and what you put into the field (fertilizer, water, fungicide, etc.) affects the output (yields) differently, depending on the characteristics of the soil and the environment (FAO, 2014; Riquelme et al., 2009). It is important for farmers to accurately identify what the needs of the field needs of for better management, which can result in the reduction of input waste and increased profit margins. While traditional agriculture has the farmer apply the same amount of input (fertilizer, water, macronutrients, pesticide, fungicide, etc.) to the entire field, which can lead to areas in the field receiving too much or too little, precision agriculture uses techniques to identify what the specific needs are in each area for targeted application, thus reducing the overall amount of the input used, saving resources, and reducing environmental impact (Santos et al., 2014). The PA cycle is depicted in Fig. 14.1.

In the PA cycle, the first step is to acquire data such as soil characteristics, weather conditions, or disease and pest information. These measurements can be done with remote sensing using active or passive sensors or more invasive techniques by soil sensors or taking samples into the lab. Often this data must be coupled with locational data, which will aid in the creation of a prescription map described in the third step. The second step is to process the data. There is vast research in this area of precision agriculture using big data analysis techniques. Much of this analysis is completed using statistical techniques or application software. The third step is to create a prescription from the analyzed data. This prescription dictates how much of an input should be applied in a specific location. This is usually in the form of a prescription map that is compatible with the applicator devices. In the fourth step, the map is integrated into the applicator devices, such as irrigation devices or fungicide applicators, which can follow the directions

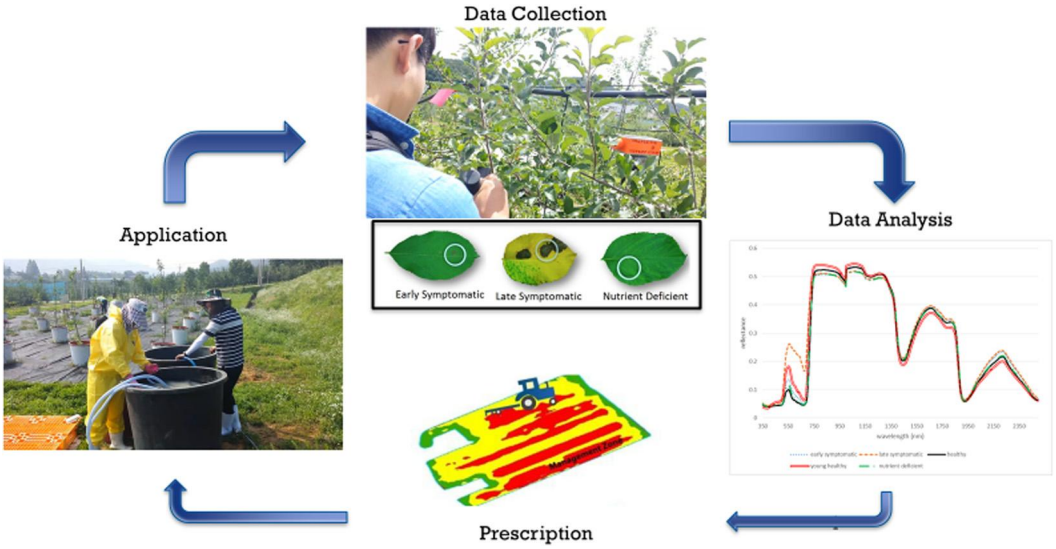


FIGURE 14.1 Flow chart of the precision agricultural process, including researchers implementing the PA process in an apple orchard in Gunwi, Republic of Korea, July 16, 2015. Courtesy of Brianna Posadas.

and vary the amount of input applied throughout the field. Once the application is complete, the precision agricultural process can begin again. The repetition of the cycle will depend on the input. For example, for irrigation the cycle will be repeated on a weekly or daily basis (Ess and Deere, 2003).

14.1.1 AI in agriculture

Machine learning algorithms have been used in a wide range of precision agricultural applications from yield prediction, disease detection, weed detection, crop quality, species recognition, animal welfare, livestock production, water management, and soil management (Liakos et al., 2018). Examples of the data types found on a farm are depicted in Fig. 14.2. Amatya et al. used colored digital images of leaves, branches, and the cherry fruits to detect cherry branches with full foliage using Bayesian models (BM)/Gaussian naive Bayes (GNB) with 89.6% accuracy (Amatya et al., 2016). Moshou et al. used spectral reflectance features to detect yellow rust infected and healthy winter wheat canopies using artificial neural networks (ANN)/multi-layer perceptron (MLP); they were able to detect yellow rust

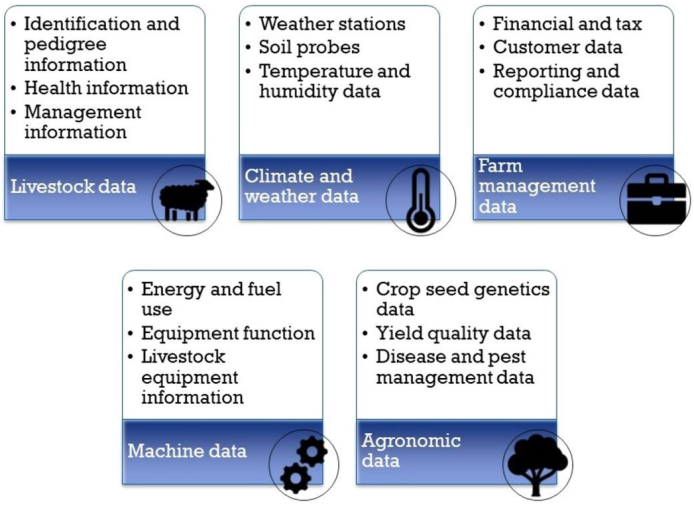


FIGURE 14.2 Different types of data found on a farm.

infected wheat with 99.4% accuracy and healthy wheat with 98.9% accuracy (Moshou et al., 2004). Pantazi used spectral bands of red, green, and NIR and texture layer to detect and map Silybum marianum using ANN/counter propagation (CP) with 98.87% accuracy (Pantazi et al., 2017). Maione et al. used twenty chemical components found in rice samples with an inductively coupled plasma mass spectrometer to predict and classify the geographical origin of a rice sample using ensemble learning (EL)/random forest (RF) with 93.83% accuracy (Maione et al., 2016). Grinblat et al. used vein leaf images of white beans, red beans, and soybean to identify and classify the 3 different species using deep learning (DL)/convolutional neural network (CNN) with over 90% accuracy (Grinblat et al., 2016). Dutta et al. used EL/bagging with tree learner to classify cattle behavior with 96% accuracy (Dutta et al., 2015). Hansen et al. used deep neural networks (DNN) for pig face recognition with 96.7% accuracy (Hansen et al., 2018). Mohammadi et al. used average air temperature, relative humidity, atmospheric pressure, vapor pressure, and horizontal global solar radiation to predict daily dew point temperature using ANN/extreme learning machine (ELM) with over 98% accuracy (Mohammadi et al., 2015). Coopersmith et al. used precipitation and potential evapotranspiration data to evaluate soil drying for

agricultural planning using instance-based models (IBM)/k-nearest neighbor (KNN) and ANN/BP with 91–94% accuracy (Coopersmith et al., 2014).

14.1.2 Big data in agriculture

To implement the machine learning algorithms for precision agriculture on a larger, industrial scale, agricultural technology providers (ATPs) collect input and output data from farmers to build prescription maps, which farmers can program farm equipment to follow (Leone, 2017). ATPs include companies such as John Deere and The Climate Corporation. For years they have been upgrading the equipment sold to farmers with sensors that passively collect information and send it back to the ATP, often without their knowledge and protected by legal contracts (Carbonell, 2016). The type of data ATPs collect include climate and weather data, agronomic data, machine data, and livestock data (Kamilaris et al., 2017). The type of ag data that is collected and used by ATPs include weather, soil types, planting, materials, spraying, yields, imagery, and other data related to farm management (Wolfert et al., 2017). A summary of agricultural data types are in Fig. 14.2. By 2025, it is predicted that an average-size farm will produce more than one million ag data points a day (Shipman, 2019). The use of PA has allowed farmers to use less resources, which saves money and reduces environmental impacts. As PA has grown as a field, and its dependence on data from hundreds of farms to create its models has increased, PA has had to rely on the techniques of big data (Mathivanan and Jayagopal, 2019; Leone, 2017). Although PA technologies generate massive amount of data, these technologies operate within a space of existing institutions, economy, governments, and politics.

14.1.3 Political economy of PA

ATPs promote PA as technologies capable of improving agricultural productivity, efficiency, and reduction in ecological footprints (Bronson, 2019; Fraser, 2019). PA technologies are fitted with sensors that collect ag big data, such as soil and animal health (Rose and Chilvers, 2018; Rotz et al., 2019). Ag big data can improve the decision-making process for farmers. However, there are concerns that the current model of agriculture through PA tech-

nologies is geared toward the industrial model and reveals a long-standing trajectory to increase the industrial model, where PA might perpetuate productivist values. These long-standing productivist values are currently driving the design of PA technologies, just like previous innovations in agriculture, such as the green and biotechnology, which created economic benefits for farmers and at the same time created power asymmetry between farmers and ATPs. The introduction of PA is indeed exacerbating power inequalities between farmers and ATPs (Bronson, 2018; Clapp and Ruder, 2020).

With inherent characteristics of technology as neither good, bad, or neutral (Kranzberg, 1986), PA is a double-edged sword that is currently designed to benefit ATPs over farmers through the generation of big data. Several concerns such as ownership, access, and power asymmetry, and privacy concerns have emerged (Carbonell, 2016; Fraser, 2019). ATPs, such as Climate Cooperation, have end-user agreements and contracts protected by intellectual property rights that prohibit farmers from modifying and repairing their farming equipment. These contracts create power asymmetries that allow ATPs to have control and access over farm data, which shifts power to the hands of ATPs (Carbonell, 2016; Wolfert et al., 2017).

The complexity and lack of transparency around end-user agreements and contracts about how big data is generated, constitute part of a larger and increasing consolidation of power and control by few corporate entities in the ag industry also raises questions about trust in farming recommendations that ATPs offer (Wiseman et al., 2019). Most ATPs do not reveal the processes involved in data collection and storage, or how the data collected are transformed into insightful farming recommendations. The entire process is essentially a "black box," where farmers have no full knowledge of how PA works (Miles, 2019). Technologies designed by ATPs monitor the everyday activities of farmers from planting to harvesting. Through this process, ATPs benefit through predicting farmers' activities to sell seeds and agrochemicals (Wolfert et al., 2017). For PA to meet the promise of crop productivity and reduction in ecological footprints from farming activities, ATPs will need to ensure that PA technologies are fair and trusted, where farmers understand and have access to decision-making for their

uses before recommendations occur. Hence some sought of assurance on the operation of these technologies.

To create trusted PA systems, ATPs will need to rely on transparency, which is the foundation for socially responsible practices in an industry increasingly controlled by few decision-makers separate from the farmstead. Addressing the lack of transparency in PA systems will require developing explainable AI (XAI) systems. The objective of XAI is to ensure that technologies can be transparent in how decisions are made and where end-user understand why specific decisions are made. In other words, XAI aims to produce explainable technologies and algorithms, where end-users understand processes that go into making recommendations. (Batarseh et al., 2021). XAI is currently a framework that explains these technologies (Wells and Bednarz, 2021). XAI ensures that the deployment of technologies is fair, transparent, and accountable (Barredo-Arrieta et al., 2019). In the next section, we discuss the current methods used for AI assurance and XAI in agriculture.

14.2 Current methods of AI assurance in agriculture

For the purposes of this chapter, we will be using Batarseh et al.'s definition of AI Assurance:

> *"A process that is applied at all stages of the AI engineering life-cycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users" (Batarseh et al., 2021)*

The study of AI Assurance is a newer development for the agricultural field. Only one paper was cited in Batarseh et al.'s (2021) review of the current state of AI Assurance: a paper on applying Data Science in agricultural policy with assurance using knowledge-based systems (Batarseh and Yang, 2017). Using this definition as a guiding principle, we have evaluated the current state of AI Assurance in agriculture by reviewing 7 studies. These studies were found using the snowball method, where inclusion criteria in-

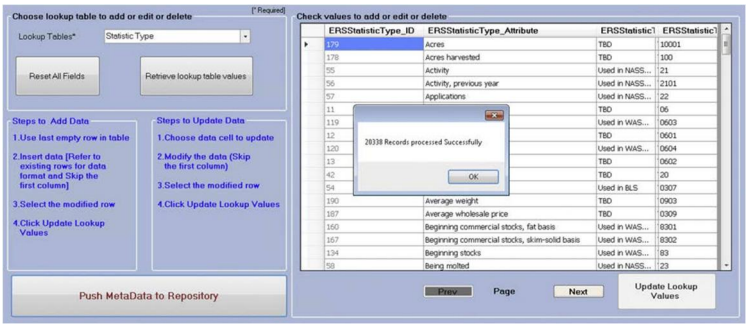


FIGURE 14.3 Screenshot of managing data streaming into the federal agricultural data warehouse through *Intelligent Federal Data Management Tool* (Batarseh et al., 2018).

cluded literature in English which specifically mentioned AI assurance, XAI, or interpretable machine learning for an agricultural application. Overall, the studies were tested by the researchers themselves or unspecified experts. However, in agriculture, we need to be cognisant that the end-users of these systems are not only the original researchers, but also the grower or agricultural worker. The wider web of stakeholders must also be taken into consideration as they are also affected by the data and algorithms that are used in modern agricultural systems. Consumers should also be factored into the AI assurance models for it to be truly transparent. Increasing the transparency of agricultural data can compel producers to improve the environmental performance of their products and close the gap between producers and consumers. Consumers will be able to trace the origins of their products and understand and monitor the system under which it was produced (Zaks and Kucharik, 2011). These recommendations will be expanded on in later sections.

14.2.1 AI assurance in agricultural policy

Batarseh et al. (2018) created a suite of engines and tools for federal teams: *Intelligent Federal Math Engine*, *Validation Engine*, and *Intelligent Federal Data Management Tool*. A screenshot of *Intelligent Federal Data Management Tool* is in Fig. 14.3. The *Intelligent Federal Math Engine* was built using *Dynamic SQL*, which aid federal analysts to calculate numerous mathematical formulas and stores the results in a database. The *Validation Engine* can

then check whether the data generated is valid. The *Validation Engine* was built using 3 SQL stored procedures *Flag validation, Technical validation,* and *Summary statistics validation*. The user interface utilizes the flag system to visually present if the data is within an expected range and is valid. The *Intelligent Federal Math Engine* is used to execute both the *Intelligent Federal Math Engine* and the *Validation Engine*, along with other services. In designing the *Intelligent Federal Data Management Tool*, researchers collected software and hardware requirements from federal teams. Key requirements were the following:

- Analysts should have the ability to perform automated validations on data (and use the validation engine).
- Analysts should be able to update all the agricultural commodity-specific metadata.
- Based on requirements, analysts should be able to change the privacy level of data. Data privacy levels could be public, private, or confidential.
- A feature to manage all the data migration rules. A knowledge base that includes mappings between two sources should be easily manageable. Analysts should be able to add, update, and delete the data migration rules (Batarseh et al., 2018).

Currently, the *Intelligent Federal Data Management Tool* is deployed at the US Department of Agriculture. Researchers surveyed federal analysts and agricultural researchers about their use of the tool, and found that 57% of the users had positive feedback on the tool.

14.2.2 AI assurance in precision agriculture

The earliest study on AI assurance in agriculture is from 2017, demonstrating just how new this field is to the agricultural realm. The differential evolution-based cooperative and competing learning of compact rule-based models, or *DECO$_3$RUM*, is a Mamdani fuzzy rule-based system for modeling problems which follows the genetic cooperative competitive learning approach using the differential evolution algorithm for its learning algorithm. An overview of *DECO$_3$RUM* is in Fig. 14.4. *DECO$_3$RUM* was tested with a local soil spectral library. Using 50 soil samples collected in

Algorithm 1 Overview of DECO$_3$RUM

Input: The E_{trn} dataset
1: Split the dataset to E_{trn} and E_{val}
2: Create the initial population of rules P_{init}
3: $P_{\text{best}} \leftarrow P_{\text{init}}$
4: **while** the termination criteria are not met **do**
5: $\quad$ **SCR** $\leftarrow$ DE(P_{best})
6: $\quad P_{\text{evolved}} \leftarrow$ FuzzyTokenCompetition(**SCR**)
7: $\quad$ **if** $\text{fit}_{\text{global}}(P_{\text{evolved}}) \geq \text{fit}_{\text{global}}(P_{\text{best}})$ **then**
8: $\quad\quad P_{\text{best}} \leftarrow P_{\text{evolved}}$
9: Simplification of P_{best}
10: Tuning of Membership Functions and rule base

FIGURE 14.4 High level overview of *DECO$_3$RUM* (Tsakiridis et al., 2017).

Central Macedonia, Greece, the system was used to predict the soil organic matter, electrical conductivity, and concentration of magnesium cations from its spectral signatures as compared to the predictions from using the partial least squares regression (PLSR) algorithm. *DECO$_3$RUM* was able to generate statistically better results as compared to PLSR. Researchers improved the *DECO$_3$RUM* system to be more amendable to hyperspectral data. These changes included using the Mahalanobis distance in the k-nn algorithm distance metric to optimize principal component space, using a feature alignment mechanism to facilitate the use of a smaller number of wavelengths in the premise part of the rules and having the predictions of the model corrected by using known errors that were discovered in the calibration set (Tsakiridis et al., 2019).

This improved system was tested on the LUCAS topsoil database, a database of 19,036 soil samples from 23 EU countries. *DECO$_3$RUM* was used to predict soil properties from the spectral signature of the samples. As compared to soil science industry standard algorithms, the system was able to either statistically outperform or produce statistically equivalent results. Though this second study mentions how the system provides "an enhanced interpretability degree for the experts" over black box models, the researchers do not define who they are referring to as "experts" (Tsakiridis et al., 2019). These researchers do build on this omission in their next study, where they explicitly design a system with an agricultural expert in mind.

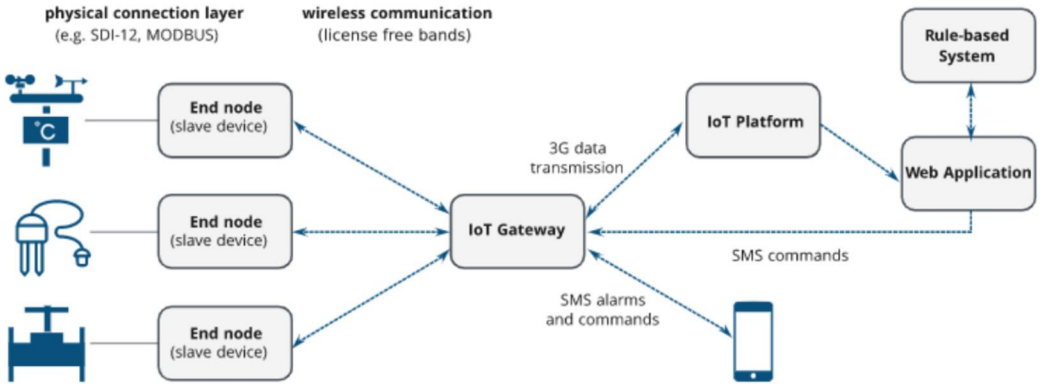


FIGURE 14.5 Overview of the Vital system (Tsakiridis et al., 2020).

With the increase use of opaque decision systems making decisions in precision agriculture, there is an increase need to understand the underlying AI mechanics by the end-users in the field. Tsakiridis et al. (2020) built an explainable AI (XAI) system called *Vital* for this purpose. It uses low-cost sensors, data store, and an explainable AI decision support system to output a fuzzy rule-base. *Vital* uses a web application as the main user interface of the XAI system, where users can visualize the inner manifestations of the rule-based system, thus making the model transparent and trustworthy (Tsakiridis et al., 2020). In one of the pilot studies, the researchers integrated their system with existing (legacy) network stations to demonstrate that *Vital* was capable of replacing aging or malfunctioning technologies. The integration of environmental sensors at Lake Koronia was successful. An additional study was conducted at the Farm School, the Aristotle University of Thessaloniki, with a precision irrigation for an olive tree orchard, where the fuzzy knowledge base was preselected by an agronomist using the linguistic description of the input variables to demonstrate that non-computer science experts could develop an AI-based model. The XAI system with the agronomist input was able to save more water than the control, irrigation manually by an agronomist (Tsakiridis et al., 2020). An overview of the *Vital* system is depicted in Fig. 14.5.

Gandhi et al. (2021) builds off the *Vital* system to build a fuzzy-rule-based system to predict the crop yield in a simulation. Using a rule-base

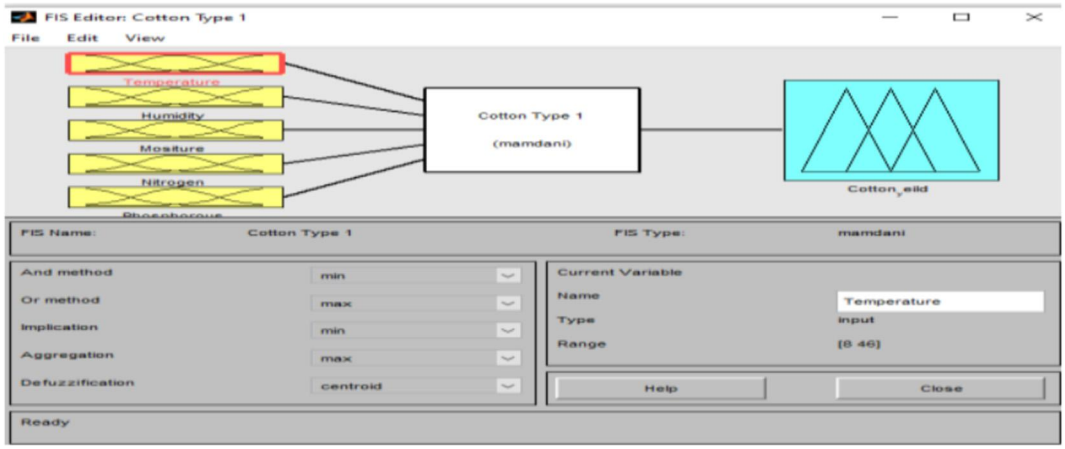


FIGURE 14.6 Screenshot of Gandhi et al. (2021)'s system for predicting crop yield.

system, the researchers aimed to simulate crop yield for cotton, wheat, paddy, barley, and maize using temperature, soil moisture, humidity, nitrogen, phosphorous, and soil type. Their models were able to generate predictions comparable to ministry of agriculture and farmers welfare. A screenshot of the system is in Fig. 14.6. Though the researchers discussed farmers using their system, the current study did not have any agricultural workers testing their system, unlike the next study in this section, where the system was built for and with agricultural workers.

Rojo et al. (2021) developed the *Augmented by Human Model Selection (AHMoSe)* system to explicitly address XAI in agriculture. The system was built to aid domain experts to better understand and compare different regression models. The system was validated by viticulture experts to aid in the selection of prediction models for grape quality. Though many XAI systems seek to make the algorithms more explainable to machine learning experts, not many systems focus on transparency and explainability for users who are experts in their domain and not necessarily ML experts (Rojo et al., 2021). The backbone of the AHMoSe system is the SHAP framework as it has been found to be more intuitive and is a good fit for users who will be comparing the system's explanations with their domain knowledge. An overview of the *AHMoSe* system is depicted in Fig. 14.7, and the user interface is in Fig. 14.8.

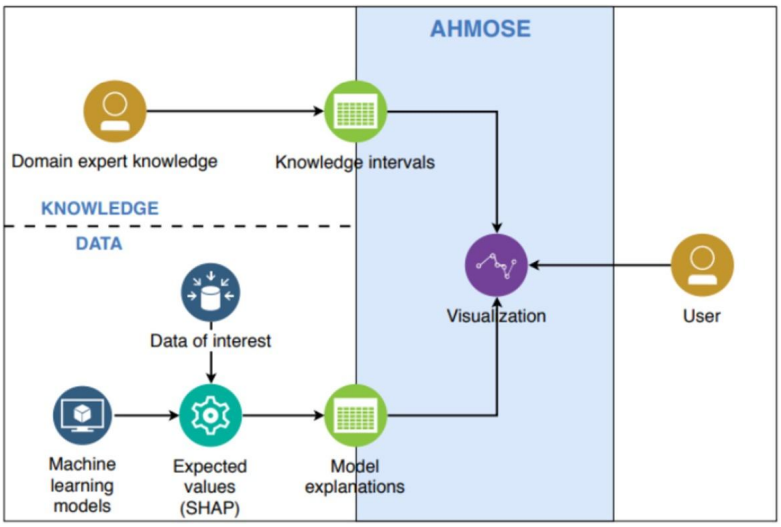


FIGURE 14.7 An overview of the flow of data in the AHMoSe ecosystem from Rojo et al. (2021).

FIGURE 14.8 The AHMoSe interface: a) the sidebar controls to select the use case, intervals, models, and features to be visualized, b) the scatter plots highlight the comparisons between models' predictions (orange and blue (gray and dark gray in print version) dots) and domain expert knowledge (green (light gray in print version) rectangles), c) the Marimekko charts indicate the importance of each feature according to the given model, and the agreement between the model and domain expert knowledge (Rojo et al., 2021).

While designing the system, researchers interviewed viticulture experts about what questions they have when choosing a machine learning model for their data:

- Should I use this model with this data?
- Why should I trust this ML model?
- Which model should I select? (Rojo et al., 2021)

Using these questions as a guide, the researchers then defined the following tasks for the AHMoSe system:

- Understand model explanations
- Identify model bias
- Compare two different model explanations
- Identify a model (Rojo et al., 2021)

The AHMoSe interface and system underwent a user evaluation with viticulture experts. Thematic analysis demonstrated 4 main themes:

- Potential use cases
- Trust
- Usability
- Understandability (Rojo et al., 2021)

Viticulture experts expressed different potential use cases for the system, including detecting anomalies in data and quality assurance. The visualizations of the system helped domain experts to trust the model and increase their likelihood of using it in their work. As much as the system inspired trust, domain experts still relied on the explanations from the researchers to understand the visualizations and the logic behind the models. As the system is now, it is not a standalone system, and improvements are needed for it to be used independently by viticulturalists. The understandability of the system could also use some improvements, with domain experts asking for a success rate of each model, instead of the root-mean-square error (RMSE), which does require some ML background to interpret (Rojo et al., 2021). A summary of the AI assurance examples in agriculture are in Table 14.1.

The current state of AI assurance in agriculture is small. While in other domains, XAI and AI assurance have been focused on making ML models understandable to ML experts, such as the Vital system. What would be more useful in agriculture is for XAI to be focused on the end users, as

Table 14.1 Summary of AI Assurance in Agriculture Studies.

Author	System	Test Conditions	End Users	Models/ Algorithms	Results
Rojo et al. (2021)	*AHMoSe*	choosing ML model to predict grape quality	viticultural-ists	knowledge-based fuzzy inference system	viticulturalists using *AHMoSe* were able to select a model with better performance than an AutoML system did
Gandhi et al. (2021)	N/A	simulating crop yield for cotton, wheat, paddy, barley, and maize using temperature, soil moisture, humidity, nitrogen, phosphorous and soil type	not specified	fuzzy-rule based system	the predicted ideal crop conditions and soil types for maximum yield were comparable to conditions provided by the Ministry of Agriculture and Farmers Welfare
Tsakiridis et al. (2020)	*Vital*	replacing legacy technologies: a set of environmental sensors in Lake Koronia	not specified	fuzzy-rule based system	integration successful
Tsakiridis et al. (2020)	*Vital*	precision irrigation for young olive tree orchard	agronomists	fuzzy-rule based system	using *Vital* with an expert was not as effective at conserving water as using *Vital* alone
Tsakiridis et al. (2019)	*DECO$_3$RUM*	LUCAS topsoil database	unspecified experts	Mamdani fuzzy rule-based system	*DECO$_3$RUM* statistically out-performed global models
Tsakiridis et al. (2017)	*DECO$_3$RUM*	to predict soil properties in samples from Central Macedonia, Greece	unspecified experts	Mamdani fuzzy rule-based System	*DECO$_3$RUM* statistically out-performed the partial least squares regression algorithm
Batarseh and Yang (2017)	*Intelligent Federal Data Management Tool*, *Intelligent Federal Math Engine*, and *Validation Engine*	suite of engines and tools at the US Department of Agriculture to manage, validate, calculate and stream data	federal analysts and agricultural researchers	knowledge-based system and data mining methods	57% of analysts gave the system positive feedback

demonstrated by AHMoSe. It is not enough to have a few select industry trained experts to understand the models; socially responsible AI assurance is about making the entire process transparent and accessible to farmers, and with their input in design. How this omission has overreaching effects beyond the immediate concerns in the field are discussed in the next section.

14.3 Agricultural policy

End-user agreements (EULAs) employed by ATPs require farmers to sign contracts to use PA technologies (Bronson, 2018; Carbonell, 2016). The purpose of these agreements is to gain farmer's trust. However, it allows farmers to relinquish their control and access to their farm data and autonomy (Fraser, 2019; Carbonell, 2016). These EULAs are not sufficient to ensure that farmers' privacy is protected, as there are tools that can foster manipulation and control by ATPs and third-party agents, who gain access to farmers' information (Wiseman et al., 2019). Weaknesses in EULAs has motivated organizations such as Ag Data Coalition in the United States to purpose that farmers play a role in how big data should be managed.

Outside of the United States, the European General Data Protection Regulation (GDPR) has set a legal framework to protect personal data similar to the federal trade commissions that oversee how data is collected in health and financial sectors. These frameworks serve to oversee the unfair practices that might be associated with data use and collection, and assign exclusive privileges to producers of data (Ferris, 2017). However, these frameworks fall short of protecting farm data, such as soil yields and weather conditions, as they are not considered personal data (Atik and Martens, 2020). The present data privacy and security legislation in the United States remain inadequate to secure and protect data generated by farmers. Hence agricultural farm data might require some level of federal and state legislation to be protected.

The lack of clear regulatory frameworks for non-personalized agricultural data is perceived as a shortcoming of state regulation by industry and private sector actors (Atik and Martens, 2020). The private sector has initialized voluntary rules and principles, in which industry actors, such as

farmers and technology innovators, need to ensure that agricultural data collection needs to follow certain codes and practices, such as consent, transparency, and disclosure, to improve how data generated from farms through PA technologies ought to be collected and utilized. These codes of practices sit alongside and complement higher government legislations (Sykuta, 2016; Sanderson et al., 2018). As a result of regulatory inadequacies, we provide suggestion on how PA might be more reassuring for farmers in the next section.

14.4 AI assurance in agriculture recommendations

Going back to the working definition of AI assurance:

> *"A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users" (Batarseh et al., 2021)*

it is clear that current practices in agriculture are not meeting this standard. Of the few studies emerging in this subfield, only a couple had explicitly designed their systems for laymen end users. And of those studies, none of them utilized industry-standard usability evaluation methods. In this section, we propose approaches and considerations for AI assurance systems in agriculture moving forward.

14.4.1 Participatory design from the start

The notion of responsible innovation (RI) is gaining ground, charting a direction for technological innovation development and how stakeholders should engage consciously and responsibly in an innovation process. Hence RI has become an essential tool for science, research, and innovation policy. RI builds on a governance and innovation assessment approach with the ambition to integrate ethical, societal values, and norms from the beginning of technology development. RI focuses on democratizing the process of innovation development built on engaging both public and

stakeholders in the process of innovation development (Owen et al., 2012; Stilgoe et al., 2013). Technology stakeholders and end-users engage in a deliberative process to innovate, while anticipating the innovation process's possible consequences. Engaging stakeholders from the early stage of innovation development can help understand the collective responsibilities that ensure that innovation is ethical, acceptable, and socially plausible (Von Schomberg, 2012). These deliberative processes can enhance innovation adoption and ensure that society adequately benefits from these innovations (Ribeiro et al., 2017; Von Schomberg, 2012), while mitigating existing societal challenges or creating new challenges. RI has been defined by Stilgoe et al. (2013) as "taking care of the future through collective stewardship of science and innovation in the present." In this way, agricultural stakeholders can work together to build technologies that can include societal values and ethics in the design of these technologies.

There is evidence to support the use of participatory design or user-centered design in agriculture to improve the use of technologies and data interpretation in the field. It has been shown that data support systems (DSS) support farmers as they strive for sustainable development. However, current DSS systems on the market fall short in fully supporting the farmer in their goals. Lindblom et al. blame this failure on the design of the DSS systems, which incorporate data that scientists thought was important for the farmer, and did not include farmer's actual needs in the design process. This failure, along with poor user interface design and their perceived problem of complexity has contributed to the "gap of relevance," which can be bridged by user-centered design (Lindblom et al., 2017). Ferrández-Pastor et al. discuss this in their study as well where they proposed a structure to include the expertise of the farmer in the design process (Ferrández-Pastor et al., 2018, 2017). An overview of the proposed process is depicted in Fig. 14.9.

Zaks and Kucharik have also discussed the importance of including end-users in the development of agroecological monitoring infrastructure. In their paper reviewing the design of these infrastructures, they noted that the poor data interpretation of the user was not because of the sensor technology itself, but the lack of integration into the agricultural management

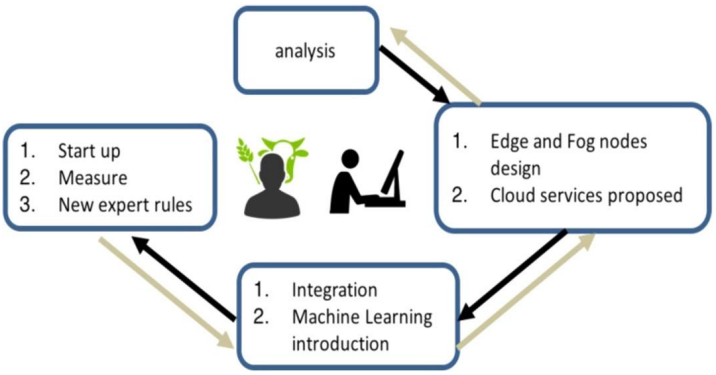


FIGURE 14.9 Design for user centered design in agriculture (Ferrández-Pastor et al., 2017).

tools end-users were familiar with, which would present the output data in a familiar format for the farmer. If the end-users could interpret the data more easily, then it would be easier to make decisions that mitigate the negative impacts of agriculture on the environment (Zaks and Kucharik, 2011). They acknowledged that to make the most robust systems possible, it will take collaboration with the end-users, policymakers, and researchers to achieve our environmental and agricultural goals with these technologies.

14.4.2 XAI for agricultural end users

Moving forward, the recommendations for AI assurance in agriculture must take into account the end user. XAI cannot only be for ML experts, but also be understandable to end users in the field (Glomsrud et al., 2019). XAI systems must also take into account the various cultural and environmental needs of the end users. As explained by Heldreth et al., trust is the most critical element for the success of AI tools (Heldreth et al., 2021). To achieve this, practitioners need to adopt these practices:

- Help users understand AI's capabilities.
- Be transparent about data and privacy.
- Recognize that many recommendations are high stakes.
- Leverage existing trusted resources (Heldreth et al., 2021).

The first practice can be achieved through the participatory design practices explained earlier in this section. Instead of only focusing on develop-

ing the algorithms, programmers need to also take into consideration how these systems will be used in the field by the end users. Addressing concerns about big data privacy and security described earlier in this chapter, developers should adopt the practice of soliciting informed consent from farmers before accessing and using their data. Developers need to be realistic about what their AI systems can actually deliver in the field; overpromising the capabilities of the systems leads to the mistrust of researchers in this industry, but is also detrimental to the end users, who ultimately pay the price for over-relaying on a system that ultimately does not live up to the abilities touted by the developer. Lastly, programmers need to work with the experts already plugged into the farming community: extension agents, input suppliers, and cooperatives. These experts have worked with the community, understand the needs, and can promote the adoption of trustworthy and transparent systems (Heldreth et al., 2021). Rural electric cooperatives are another trusted organization that can assist in bringing connected agriculture technologies to farmers. This is especially important for farmers in minority or under-served communities, where the advantages of connected agriculture can have an even higher payoff. Adopting the framework of "connected ag as a service" (CAAAS) can also help small farmers take advantage of the capabilities of data through data cooperatives. These can be further leveraged to support and promote the use of data in ways that more directly support the local community (Rayi et al., 2021).

To develop more farmer-centered AI assurance systems, Heldreth et al. also describes barriers that practitioners need to be aware of:

- Common hardware and data constraints
- Build for diverse literacies and multiple languages
- Co-design with smallholder farmers and intermediaries (Heldreth et al., 2021)

Peters et al. discusses the need for more human-guided ML to be integrated into AI technologies in agriculture (Peters et al., 2020). Glomsrud et al. recommend that different explainable AI models should be developed for different users:

- Developer
- Assurance

- End-user
- External (Glomsrud et al., 2019)

In agriculture, the developer and assurance users are typically researchers at ATPs or universities. In general, AI assurance models already serve these users. More AI assurance systems are needed, which focus their service on the agricultural end-user, who can be farmers, ranchers, agronomists, viticulturists, soil scientists, or citizen scientists. Examples of external users in agriculture include policymakers and consumers. Each of these users will have different data needs and have different preference for XAI user interfaces. Studies in AI assurance in agriculture need to be explicit about which user they are building for and include these users in the design. Following in the model of Rojo et al., XAI systems developed for agriculture should also be evaluated for usability using user-centered design principles to further ensure these systems are meeting the needs of end-users (Rojo et al., 2021).

In designing for each of these user types, Alikhademi et al. (2021) have developed a rubric for XAI systems:

- Does the XAI tool clearly identify its target audience and their expectations for the tool?
- Is the presentation of explanations sufficient for the target audience to gain insight and improve upon their model?
- Does the XAI tool provide a variety of types of explanations? (Alikhademi et al., 2021)

Addressing these questions will not only ensure the system is understandable for the specified user, but will also aid in the usability of the system. Batarseh et al. (2018) have also developed a list of usability recommendations as well, based on their own study of an AI assurance system for federal analysts. Their best practices include

- User access roles
- Data entry validation
- Multithreading
- Application integration (Batarseh et al., 2018)

Different users will need different levels of access to the data, and these roles need to be clearly defined and assigned to the appropriate users. There needs to be a system to ensure that only technical federal users are imputing data and nontechnical federal users are restricted from making changes to the data. Because federal analysts' work involves running multiple processes at once, there needs to be a capacity to execute multithreading. Lastly, the system needs to integrate well with the other applications heavily used by federal analysts (Microsoft Excel, Outlook, R, and internet browsers) (Batarseh et al., 2018).

14.5 Conclusion

The improvement of AI assurance in agriculture will not only impact the end-users who operate in the agriculture field, but also the policymakers who rely on transparent, trustworthy, and valid data to complete their work. With a focus on developing AI assurance models with the end-user in mind using participatory design and user-centered design principles, computer scientists can vastly improve the quality of work in the agricultural field. By working within communities, researchers can build to their immediate needs and create more trustworthy and transparent systems. This will allow for more socially responsible precision agriculture that responds to the needs of stakeholders and works for them. This change has rippling effects throughout the rest of the stakeholders, including consumers of agricultural products. For these studies to be effective, interdisciplinary teams will be needed and experts from social science, AI, human-computer interaction, agronomy, precision agriculture, and policy will need to collaborate to develop robust systems that meet the needs of the field.

CRediT authorship contribution statement

Brianna B. Posadas: Conceptualization of this study, Data curation, Writing – Original draft preparation. **Ayorinde Ogunyiola:** Conceptualization of this study, Data curation, Writing – Original draft preparation. **Kim Niewolny:** Writing – Original draft preparation.

References

Alikhademi, K., Richardson, B., Drobina, E., Gilbert, J.E., 2021. Can explainable AI explain unfairness? A framework for evaluating explainable AI. arXiv preprint. arXiv: 2106.07483.

Amatya, S., Karkee, M., Gongal, A., Zhang, Q., Whiting, M.D., 2016. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. Biosystems Engineering 146, 3–15.

Atik, C., Martens, B., 2020. Competition problems and governance of non-personal agricultural machine data: Comparing voluntary initiatives in the US and EU.

Barredo-Arrieta, A., Laña, I., Del Ser, J., 2019. What lies beneath: a note on the explainability of black-box machine learning models for road traffic forecasting. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2232–2237.

Batarseh, F.A., Freeman, L., Huang, C-H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8 (1), 1–30.

Batarseh, F.A., Ramamoorthy, G., Dashora, M., Yang, R., 2018. Intelligent automation tools and software engines for managing federal agricultural data. In: Federal Data Science. Elsevier, pp. 193–210.

Batarseh, F.A., Yang, R., 2017. Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence. Academic Press.

Bronson, K., 2018. Smart farming: including rights holders for responsible agricultural innovation. Technology Innovation Management Review 8 (2), 7–14.

Bronson, K., 2019. Looking through a responsible innovation lens at uneven engagements with digital farming. NJAS-Wageningen Journal of Life Sciences 90, 100294.

Carbonell, I., 2016. The ethics of big data in big agriculture. Internet Policy Review 5.

Chaterji, S., DeLay, N., Evans, J., Mosier, N., Engel, B., Buckmaster, D., Chandra, R., 2020. Artificial intelligence for digital agriculture at scale: techniques, policies, and challenges. arXiv preprint. arXiv:2001.09786.

Clapp, J., Ruder, S-L., 2020. Precision technologies for agriculture: digital farming, gene-edited crops, and the politics of sustainability. Global Environmental Politics 20 (3), 49–69.

Coopersmith, E.J., Minsker, B.S., Wenzel, C.E., Gilmore, B.J., 2014. Machine learning assessments of soil drying for agricultural planning. Computers and Electronics in Agriculture 104, 93–104.

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., Henry, D., 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. Computers and Electronics in Agriculture 111, 18–28.

Ess, D., Deere, J., 2003. The Precision Farming Guide for Agriculturists. Cengage Learning.

FAO, 2014. The State of Food and Agriculture 2014, in Brief. Food and Agriculture Organization of the United Nations Rome.

Ferrández-Pastor, F-J., García-Chamizo, J-M., Hidalgo, M.N., Mora-Martínez, J., 2017. User-centered design of agriculture automation systems using internet of things paradigm. In: International Conference on Ubiquitous Computing and Ambient Intelligence, pp. 56–66.

Ferrández-Pastor, F.J., García-Chamizo, J.M., Nieto-Hidalgo, M., Mora-Martínez, J., 2018. Precision agriculture design method using a distributed computing architecture on internet of things context. Sensors 18 (6), 1731.
Ferris, J.L., 2017. Data privacy and protection in the agriculture industry: is federal regulation necessary. Minnesota Journal of Law, Science & Technology 18, 309.
Fraser, A., 2019. Land grab/data grab: precision agriculture and its new horizons. The Journal of Peasant Studies 46 (5), 893–912.
Gandhi, R., Bhardwaj, S., Sehgal, B., Gupta, D., 2021. An explainable AI Approach for Agriculture Using IoT. Available at SSRN 3834259.
Glomsrud, J.A., Ødegårdstuen, A., Clair, A.L.S., Smogeli, Ø., 2019. Trustworthy versus explainable AI in autonomous vessels. In: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC), pp. 37–47.
Grinblat, G.L., Uzal, L.C., Larese, M.G., Granitto, P.M., 2016. Deep learning for plant identification using vein morphological patterns. Computers and Electronics in Agriculture 127, 418–424.
Hansen, M.F., Smith, M.L., Smith, L.N., Salter, M.G., Baxter, E.M., Farish, M., Grieve, B., 2018. Towards on-farm pig face recognition using convolutional neural networks. Computers in Industry 98, 145–152.
Heldreth, C., Akrong, D., Holbrook, J., Su, N.M., 2021. What does AI mean for smallholder farmers? A proposal for farmer-centered AI research. Interactions 28 (4), 56–60.
Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture 143, 23–37.
Kranzberg, M., 1986. Technology and history: "Kranzberg's laws". Technology and Culture 27 (3), 544–560.
Leone, L., 2017. Addressing big data in EU and US agriculture: a legal focus. European Food and Feed Law Review 12, 507.
Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors 18 (8), 2674.
Lindblom, J., Lundström, C., Ljung, M., Jonsson, A., 2017. Promoting sustainable intensification in precision agriculture: review of decision support systems development and strategies. Precision Agriculture 18 (3), 309–331.
Maione, C., Batista, B.L., Campiglia, A.D., Barbosa Jr, F., Barbosa, R.M., 2016. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Computers and Electronics in Agriculture 121, 101–107.
Mathivanan, S., Jayagopal, P., 2019. A big data virtualization role in agriculture: a comprehensive review. Walailak Journal of Science and Technology (WJST) 16 (2), 55–70.
Miles, C., 2019. The combine will tell the truth: on precision agriculture and algorithmic rationality. Big Data & Society 6 (1), 2053951719849444.
Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., Gocic, M., 2015. Extreme learning machine based prediction of daily dew point temperature. Computers and Electronics in Agriculture 117, 214–225.
Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., Ramon, H., 2004. Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks. Computers and Electronics in Agriculture 44 (3), 173–188.

Owen, R., Macnaghten, P., Stilgoe, J., 2012. Responsible research and innovation: from science in society to science for society, with society. Science and Public Policy 39 (6), 751–760.

Pantazi, X.E., Tamouridou, A.A., Alexandridis, T., Lagopodi, A.L., Kashefi, J., Moshou, D., 2017. Evaluation of hierarchical self-organising maps for weed mapping using UAS multispectral imagery. Computers and Electronics in Agriculture 139, 224–230.

Peters, D.P., Rivers, A., Hatfield, J.L., Lemay, D.G., Liu, S., Basso, B., 2020. Harnessing AI to transform agriculture and inform agricultural research. IT Professional 22 (3), 16–21.

Rayi, P.S., Draper, Z., II, L.W., Laskey, K., 2021. Commonwealth of Virginia - Connected Agriculture: Innovation and Opportunities an Initial Research Effort. Virginia Department of Agriculture and Consumer Services.

Ribeiro, B.E., Smith, R.D., Millar, K., 2017. A mobilising concept? Unpacking academic representations of responsible research and innovation. Science and Engineering Ethics 23 (1), 81–103.

Riquelme, J.L., Soto, F., Suardíaz, J., Sánchez, P., Iborra, A., Vera, J., 2009. Wireless sensor networks for precision horticulture in Southern Spain. Computers and Electronics in Agriculture 68 (1), 25–35.

Rojo, D., Htun, N.N., Parra, D., De Croon, R., Verbert, K., 2021. AHMoSe: a knowledge-based visual support system for selecting regression machine learning models. Computers and Electronics in Agriculture 187, 106183.

Rose, D.C., Chilvers, J., 2018. Agriculture 4.0: broadening responsible innovation in an era of smart farming. Frontiers in Sustainable Food Systems 2, 87.

Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., et al., 2019. The politics of digital agricultural technologies: a preliminary review. Sociologia Ruralis 59 (2), 203–229.

Ryan, M., 2019. Agricultural big data analytics and the ethics of power. Journal of Agricultural and Environmental Ethics 1 (21).

Sanderson, J., Wiseman, L., Poncini, S., 2018. What's behind the ag-data logo? An examination of voluntary agricultural data codes of practice. International Journal of Rural Law and Policy 1, 1–20.

Santos, I.M., Da Costa, F.G., Cugnasca, C.E., Ueyama, J., 2014. Computational simulation of wireless sensor networks for pesticide drift control. Precision Agriculture 15 (3), 290–303.

Shipman, M., 2019. Interpreting the growing field of on-farm data. Retrieved from: https://news.ncsu.edu/2019/10/on-farm-data/.

Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Research Policy 42 (9), 1568–1580.

Sykuta, M.E., 2016. Big data in agriculture: property rights, privacy and competition in ag data services. International Food and Agribusiness Management Review 19 (1030-2016-83141), 57–74.

Tsakiridis, N.L., Diamantopoulos, T., Symeonidis, A.L., Theocharis, J.B., Iossifides, A., Chatzimisios, P., et al., 2020. Versatile internet of things for agriculture: an eXplainable AI approach. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 180–191.

Tsakiridis, N.L., Theocharis, J.B., Panagos, P., Zalidis, G.C., 2019. An evolutionary fuzzy rule-based system applied to the prediction of soil organic carbon from soil spectral libraries. Applied Soft Computing 81, 105504.
Tsakiridis, N.L., Theocharis, J.B., Zalidis, G.C., 2017. A fuzzy rule-based system utilizing differential evolution with an application in vis-NIR soil spectroscopy. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–7.
Von Schomberg, R., 2012. Prospects for technology assessment in a framework of responsible research and innovation. In: Technikfolgen abschätzen lehren. Springer, pp. 39–61.
Wells, L., Bednarz, T., 2021. Explainable AI and reinforcement learning—a systematic review of current approaches and trends. Frontiers in Artificial Intelligence 4, 48.
Wiseman, L., Sanderson, J., Zhang, A., Jakku, E., 2019. Farmers and their data: an examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. NJAS-Wageningen Journal of Life Sciences 90, 100301.
Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M-J., 2017. Big data in smart farming–a review. Agricultural Systems 153, 69–80.
Zaks, D.P., Kucharik, C.J., 2011. Data and monitoring needs for a more ecological agriculture. Environmental Research Letters 6 (1), 014017.

This page intentionally left blank

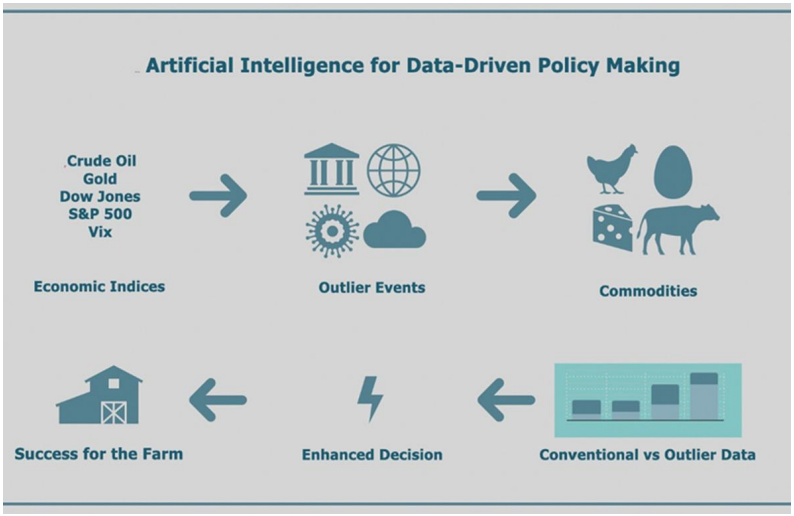
15

The application of artificial intelligence assurance in precision farming and agricultural economics

Madison J. Williams[a], Md Nazmul Kabir Sikder[b], Pei Wang[c], Nitish Gorentala[d], Sai Gurrapu[e], and Feras A. Batarseh[f]

[a]*University of Mary Washington, Fredericksburg, VA, United States* [b]*Bradley Department of Electrical and Computer Engineering (ECE), Virginia Tech, Arlington, VA, United States* [c]*Microsoft Corporation, Redmond, WA, United States* [d]*Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, United States* [e]*Apple Inc., Cupertino, CA, United States* [f]*Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States*

Graphical abstract



AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00029-9

Abstract

Agricultural policy has traditionally been conducted in an ad hoc manner, generally, by responding to natural disasters instinctively or managing unavoidable causes through the implementation of short term financial compensations or long-term loan and insurance programs at farms. The presented model, namely AI2Farm, provides farmers with predictions during conventional and unconventional times to compensate for the little to no guidance that policies, such as the Farm Bill, provide for spur-of-the-moment decisions needed to be made that arise from outlier events.

Farmers are an integral part of our economy due to their ability to manage market supply and demand expectations that solidify the nation's food security. Therefore it's important that farmers have access to the most up-to-date technology to make sound decisions that are in the best interest of rural and urban communities. Machine learning (ML) models measure associations, correlations, and causations of global and domestic events via commonplace financial indices with the production, consumption, and pricing of global agricultural commodities in the United States. Consequently, a deeper understanding of changes in behavior displayed by farmers as a result of outlier events aid in the ability to determine how precision agriculture can best assist farmers in the decision-making process. This entire set of information is lastly applied to the analysis of farms in the state of Virginia with smart tools and equipment that can benefit from models such as AI2Farm; the model and its results are presented and discussed.

Keywords

Outlier detection, economic indices, precision agriculture, smart farms

Highlights

- Artificial Intelligence (AI) systems are deployed to enable precision farming activities
- Data from economic indices such as the Chicago Board Options Exchange's CBOE Volatility Index (VIX), Gold, Oil, S&P 500, Dow Jones (DOWIA), as well as commodity data from the United States Department of Agriculture (USDA) are used
- Conventional and outlier-based predictions are presented as two alternative scenarios, where the farmer can choose from both scenarios depending on their current context
- Outlier events considered include: political, financial, environmental, health-related, global, and domestic events

- While most precision agriculture tools present localized recommendations that are disconnected from the world's state-of-affairs, the presented method provides a conventional recommendation, as well as one depending on events and their effect on farming

15.1 Introduction

The field of Precision Agriculture uses a variety of technologies, such as sensing, information technologies, and mechanical systems, to manage different parts of a field separately (USDA, 2018). The act of adopting such practice and applying it to day-to-day farm procedures is known as Precision Farming. Precision Farming provides some sense of stability amidst conditions such as weather and market demands that are natural actors within agriculture at the local and global level; protecting one's commodities and maximizing economic yield in the long run. Although farmers grow accustom to such conditions, there are instances where outlier events occur that overwhelm current monitoring and forecasting tools; prohibiting farmers from making sound decisions.

Formal acknowledgment of economic fluctuations are not merely enough to form an understanding as to *how* and *why* the extremities of outlier events vary and occur. What is required, rather, for precision agriculture, is the intersection of policy and economics to enable data scientists and public policymakers to make more informed decisions.

It is known that political events directly or indirectly affect the economy and VIX (Shaikh, 2019). COVID-19, which began at the end of 2019, is an outlier event resulting in vast disruption on the United States economy and financial markets; all of which was unforeseeable for many (Brown et al., 2021). The United States is a country that values individualism over collectivism; one where individuals are reluctant to participate in a cause if it's an inconvenience or burden to themselves despite the protection that it may provide to their neighbors (Vandello and Cohen, 1999). In turn, the notion of individualism further exacerbated the issue of COVID-19.

The manner in which the formal announcement of COVID-19 within the United States unfolded left little room for any current intelligent algorithm

to be of use. Food insecurity became an immediate concern and reality for families across the nation. Consumer consumption increased as states were advised to go into lockdown, which strained retailers across the nation. The relationship between agriculture and this particular outlier event will be a reoccurring example throughout this chapter, because, for many, this obscure event is the most relevant and well known outlier in recent memory.

This research is concerned with the effects of an outlier event, such as the pandemic, on the commodities of animal products within the United States. The pandemic is multifaceted in the sense that it can be categorized as a global and political event, which had a direct impact on the production of goods. The distribution of vaccinations, a more recent development to offset the spread of the coronavirus, for example, had a direct relation to the health of farms. The Purdue food and agriculture vulnerability index estimates nationwide, "over 496,000 agriculture workers have tested positive for coronavirus, with over 3000 in New York State alone" (Purdue, 2020). The management of their fields and crop production was jeopardized alongside their health.

The Purdue food and agriculture vulnerability index in collaboration with Microsoft served as a baseline in terms of establishing the scholarly work that is already available, and identifying what can be improved upon. Purdue University "combined data on the number of Covid-19 cases in each U.S. county with the county's total population, the U.S. Department of Agriculture data on the number of farmers and hired farm workers in each county, data on agricultural production of each county, and lastly was able to estimate the share of agricultural production at risk" (Purdue, 2020). The visualization of loss of production within various states was useful in developing a deeper understanding of the struggles within the agriculture industry, more specifically during an outlier event. Though the loss of production impact for a given commodity is an aspect of agriculture research, it's unable to be useful for prediction of the other outlier events considered in this work as well as their relationship with economic indices. In this sense, we're able to distinguish this research from Purdue University and other existing scholarly work.

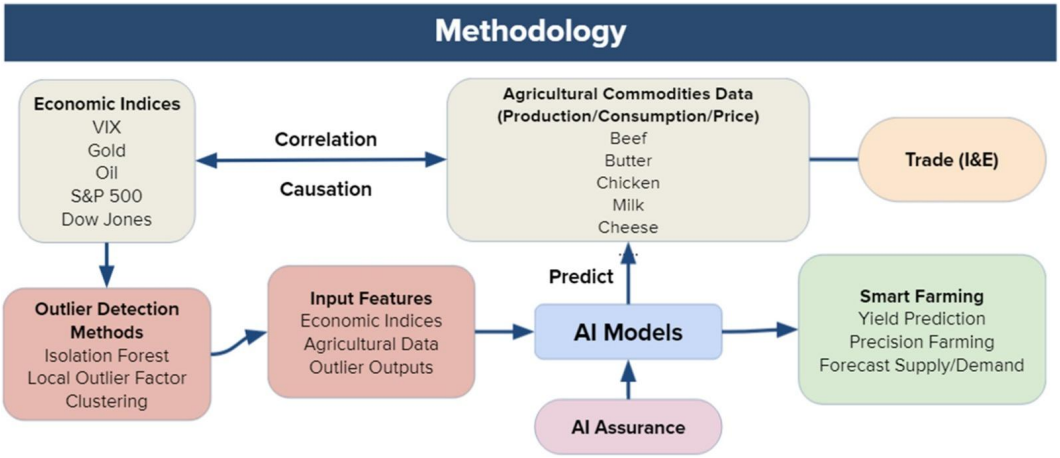


FIGURE 15.1 The AI2Farm method.

15.2 AI for smart farms

Beginning on the far right corner and continuing left in Fig. 15.1 of the AI2Farm model, the commodity of interest is observed alongside the six chosen economic indices to understand their relationship when fluctuation occurs. An outlier detection technique develops through the use of ML algorithms (i.e., isolation forest, autoencoder, K-means clustering, support vector machine) to detect unusual/outlier points on VIX, S&P500, gold, DOWIA, and crude oil indices. By properly labeling outlier data points, we're able to predict future outliers as well from the models. Definitions of outlier events for mentioned data are any anomalous dataset that behaves abnormally among the rest of the population, which in turn indicates particular events in the real world that cause the datasets to behave abnormally. Successful detection of outlier events deepens one's understanding of the effect of social/political impacts on smart farms. Both supervised and unsupervised learning models are used when comparing each of the model's performance matrices.

As it relates to the desire to centrally focus on AI for econ and international outlier events, for the purposes of technology and science policy, we're inclined to ask the following:

(a) *can policy scenarios be built to validate and optimize outcomes of different data-driven policies?*
(b) *can an economic causality model define the causes and effects of global outlier events using learning (from economic indices) and assured AI?*
(c) *can AI models factor outlier events into economic predictions to support farming decisions differently during outlier events vs. "conventional times"?*

The three questions above differentiate our research from current standing studies that merely focus on COVID-19 or another singular outlier event for that manner (Gruère and Brooks, 2021; Elleby et al., 2020). It is here we shift our focus towards weather; an outlier event that has been documented for centuries. Weather conditions, such as temperature, rain, humidity, moisture, and wind speed all impact yield production. Although documentation of such conditions through the USDA weather archives remain in use by farmers, rise in temperature caused by global warming will result in more persistent weather anomalies that will increase the need of better weather forecasting and question the use of traditional farming methods (D'Agostino and Schlenker, 2016). The urgency to implement a new form of predictive modeling pertains only to weather, just one outlier of many, all the while farmers are still subject to the impacts of supply and demand and market prices. This scenario illustrates that focusing on just one outlier event is not enough, because, in reality, farmers have to explore multiple avenues to make the best decision for their commodities. The aforementioned is why programs such as the USDA's Natural Resources Conservation Service and the National Water and Climate Center (NWCC), as traditional as it may be, might be losing their value (D'Agostino and Schlenker, 2016). Solely remaining responsible for producing and disseminating accurate and reliable forecasts and other climatic datum are trivial if it isn't specific towards a particular commodity or does not address other worldly events. Additionally, the manner in which they are collected and distributed to farmers is not of use. Generating forecasts in near real-time is the desired result of new and up incoming models, such as the AI2Farm model that is presented in this chapter.

The cost of production forecasts for US major field crops, for instance, is centered upon projecting net returns at the national level. The projected costs are based upon the previous year's production costs and projected changes in the coming year's indexes of prices paid for farm inputs (Knoema, 2021). Although the long-term baseline projections are in a sense reliable, they fail to provide an explanation as to why the fluctuation in prices occurred to begin with. The inclusion of economic indices is a starting point, but the lack of awareness surrounding outlier events and their impacts is lacking, and an area in which this research expounds. Conversely, research by the International Production Assessment Division (IPAD) of the USDA's Foreign Agricultural Service (FAS) does take outlier events into consideration. The primary mission of IPAD is to produce the most objective and accurate assessment of global agricultural production and the conditions affecting food security around the world (Becker-Reshef et al., 2010). Outcomes of the IPAD are monthly crop production estimates and early warnings of crop disasters. Though the outcome is similar in terms of early warning of crop disaster through the detection of weather outlier events, the method of collecting data is different in that it only focuses on one outlier event (weather) and doesn't concern economic indices. What continues to distinguish this research is that we have identified a new relationship that has not been studied before; considering the impact of both economic indices and outlier events. Observing one without the other is what separates research pertaining to localized versus global knowledge.

15.2.1 Correlation of economic indices and various commodities

The creation of the AI2Farm model begins with identifying the relationship between commodity production and the six economic indices. When evaluating economic indices and commodity data, it's imperative to understand the relationship between the two variables to determine whether or not correlation should be the basis of the decision-making process for a farmer. With the statistical knowledge that correlation does not equal causation, we run each causal and correlation value instead of one per production dataset. The process entails evaluating the highest causation, running

the model, and then evaluating the highest correlation and running the model again to see which one has the strongest fit of the data. The Pearson's correlation coefficient is as follows:

$$r = \frac{\sum (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \tag{15.1}$$

Pearson's correlation coefficient measures the following: if its p-value is less than the α setting (typically .05), we deem there to be a meaningful association, and the r value tells us whether the correlation is positive or negative.

As indicated in Fig. 15.2, the highest correlation goes to 1 and the lowest to 0; positive correlation is positive 1 at the highest point and the negative correlation is at -1 at the lowest point. For negative one as VIX goes up, beef will go down (exactly its opposite).

As indicated within Fig. 15.3, the x axis is the commodity, and the y axis are the indices. This is another form of representing how data are affecting one another. The correlation between the indices and the fluctuation of the commodities is indicated by either a strong positive linear correlation, as with cheese, or a nonlinear negative correlation as with beef.

15.2.2 Causation of economic indices and various commodities

Determining causality (a.k.a. causation), is the next logical sequence. In doing so, we hope to identify the causal score for the impact of each index on the production to inform farmers to use causation when possible, but when the strength is weak, to defer to correlation to determine which economic index they would like to focus on for each production.

DoWhy, an open-source Python library, is utilized to address the causal question in this research. DoWhy is unique in its ability to expand upon causal inference estimation methods, such as Python and R, that test statistical significance without the confirmation that the foundation in which it acted upon is in fact solid. Essentially, DoWhy minimizes the expectation of an analyst to not only provide their own causal model and checks for assumptions, but to provide it correctly in a manner fit for causal inference.

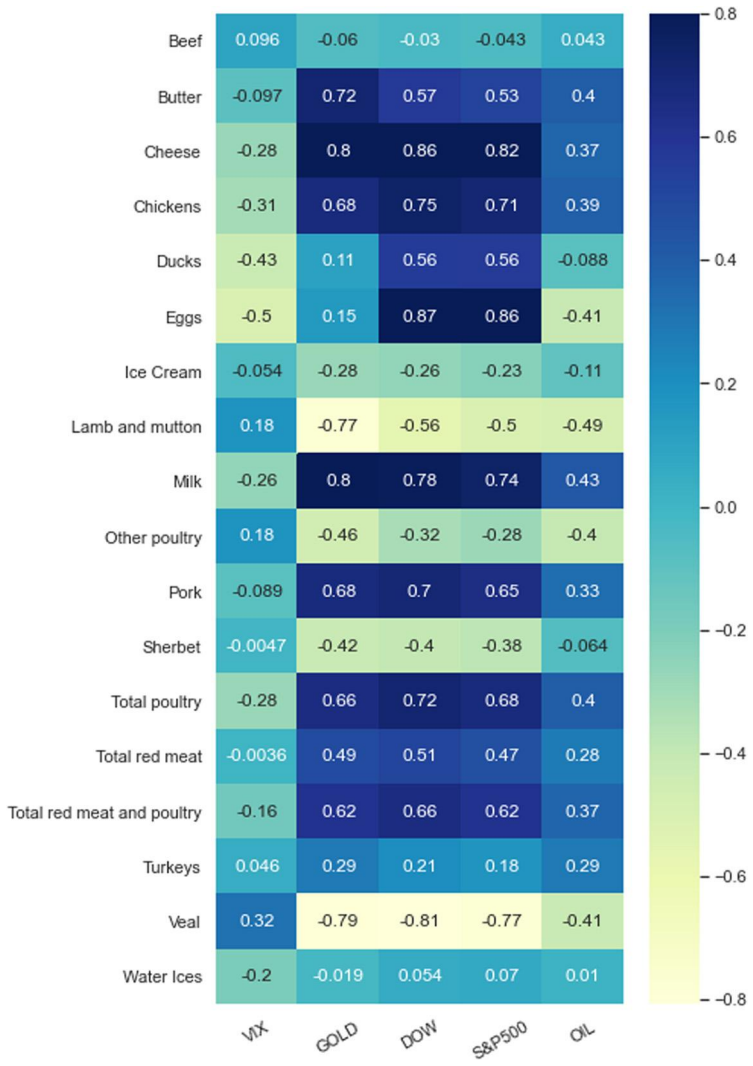


FIGURE 15.2 Correlation model between economic indices and commodities.

To relieve this burden, and to ensure that the steps prior to the estimation step were done correctly, DoWhy added an additional three steps to their pipeline, as shown in Fig. 15.4.

The four-step analysis pipeline includes the following: model, identify, estimate, and refute. Model causal mechanisms, identify the target estimated, estimate causal effect, and refute the estimate. This end-to-end li-

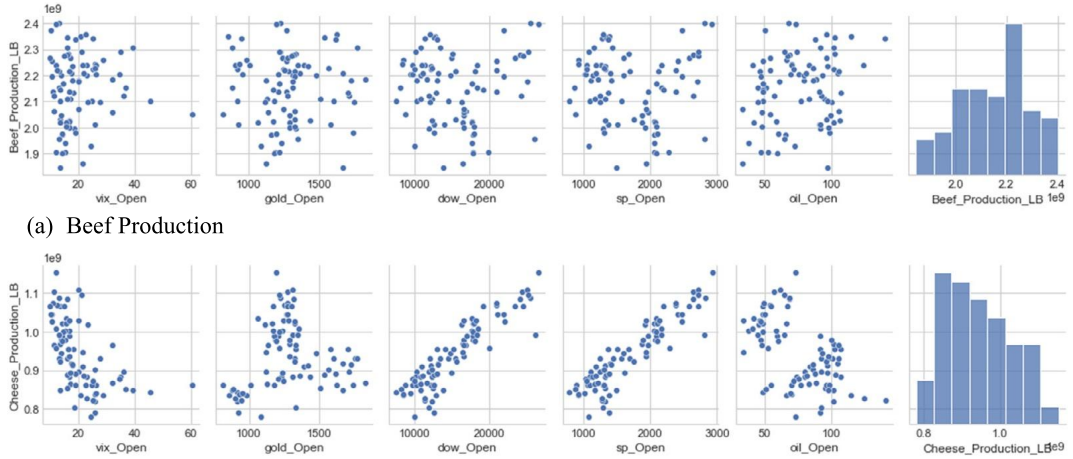


FIGURE 15.3 Commodity Production pair plot for (a) beef and (b) cheese.

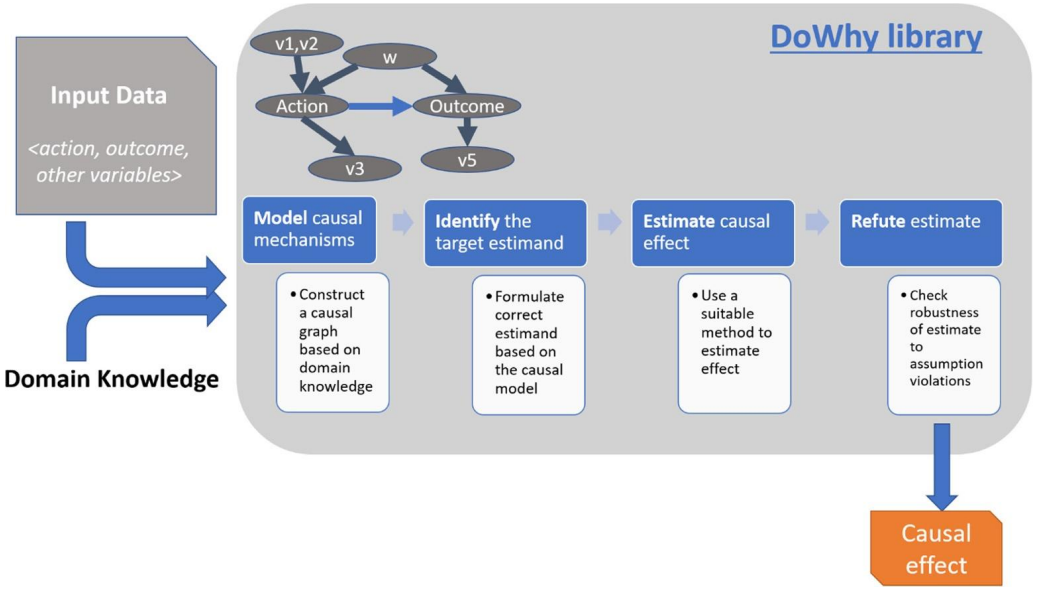


FIGURE 15.4 The four-step analysis pipeline in DoWhy (Sharma and Kiciman, 2020).

brary for causal inference provided the certainty necessary to estimate the causal effect.

Table 15.1 includes the results gathered using DoWhy. Shown above are five columns of the commodities and economic indices and eighteen rows

Table 15.1 Scores processed by DoWhy for causation between each commodity and economic indicators.

	Crude Oil Open	Gold Open	DOWIA Open	S&P500 Open	VIX Open
Beef	0.226021	−0.23521	1.608905	−1.42056	0.28295
Butter	−0.01532	0.477177	0.154803	0.145163	0.090398
Cheese	−0.05527	0.406332	0.239261	0.367241	0.110741
Chickens	0.10447	0.165347	0.305438	0.169931	0.044098
Ducks	−0.02345	−0.15335	−0.53592	1.017804	−0.21246
Eggs	−0.04621	0.043418	0.544103	0.186939	0.020712
Ice cream	0.02511	−0.12444	0.209716	−0.46338	−0.24881
Lamb and mutton	−0.0291	−0.45765	0.293928	−0.43749	0.064549
Milk	0.017964	0.4296	0.0652	0.43674	0.0775
Other poultry	−0.20287	−0.14632	−0.16872	0.058221	0.179605
Pork	0.015955	0.235899	0.84366	−0.28956	0.275482
Sherbet	0.271992	−0.40333	0.401844	−0.60832	−0.20088
Total poultry	0.121128	0.140361	0.444497	0.037026	0.079839
Total red meat and poultry	0.120495	0.089676	0.892369	−0.41871	0.192183
Total red meat	0.119753	0.030156	1.418313	−0.95388	0.32411
Turkeys	0.28781	−0.06445	0.502205	−0.25064	0.291053
Veal	0.024031	−0.47032	0.558691	−1.09531	0.070137
Water ices	0.037014	−0.04263	0.23056	−0.189	−0.2337

of the production datasets (crop/animal data). Linear regression is a statistical process to model the relationship between two variable; a method used within the third step of the pipeline to estimate the causal effect. Through the use of linear regression, DoWhy isolates one independent variable from the other independent variables to observe the effect of one thing, whilst ignoring the effects of others.

In the causal model, the arrows are reflective of indices and their "associated" production. The thought process behind using causation is to isolate one index and its effect on individual commodities: in this instance, beef. With the awareness that all five indices may have an effect on the commodity of choice, isolating all of the independent variables that we do not have control of results in more control groups. The result is a better understanding of which economic index is best to focus on for measuring production,

and therefore helping with decisions on the farm. The data normalization formula used is as follows (Loukas, 2021):

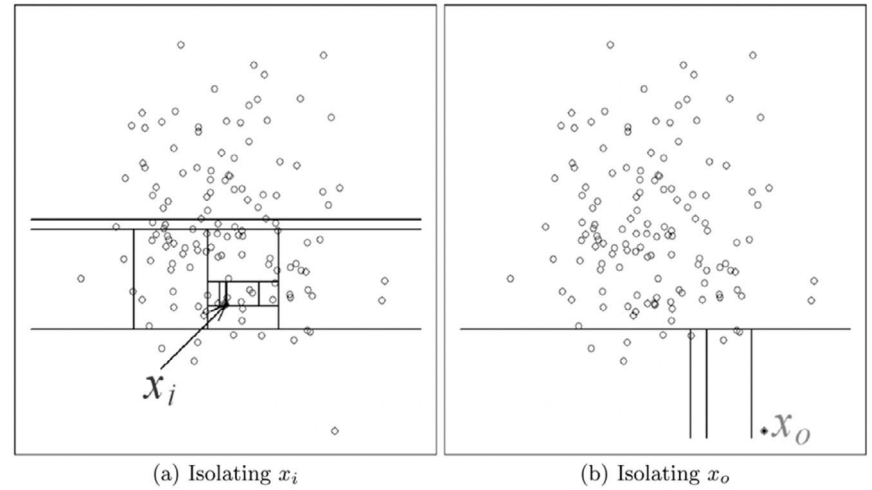
$$X\ scaled = x - \frac{xmin}{xmax} - xmin \tag{15.2}$$

The min-max scaler is used to normalize all stock price data in the range of 0 to 1 for each stock. Without this structure, it's a comparison between a large and complex index to a smaller index, which results in data bias issues further down the line.

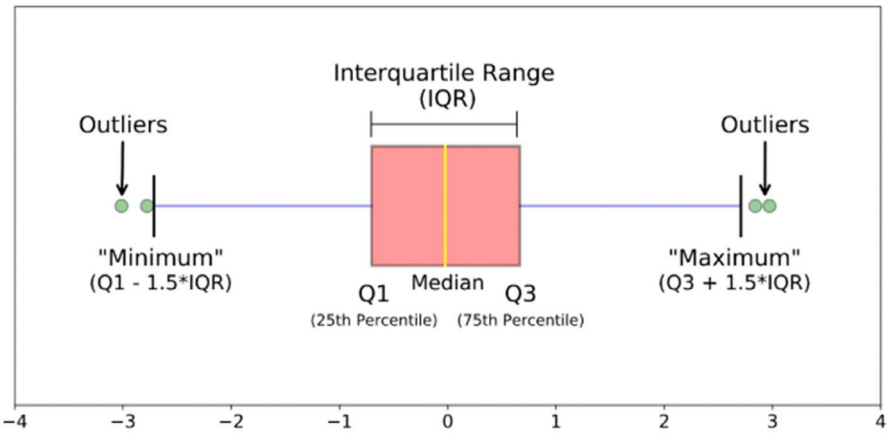
15.2.3 Scoring outlier events for the model and finding anomalies

The AI application begins with the introduction of outlier events; one from a generated algorithm and another from real world events. Both are chosen to ensure that the identification of outlier events isn't limited to the scope of the generated data, rather it is to the scope of that data and beyond with the real world events (a comprehensive list is manually collected). The unsupervised algorithm of choice is isolation forest or iForest to detect abnormal behavior within the economic dataset. This model-based method is the preference over existing distance- and density-based methods due to its ability to handle larger datasets and identify anomalies at a quicker rate. The concept of an isolation forest is as follows (Liu et al., 2012) (see Fig. 15.5):

> *"In a data induced random tree, partitioning of instances are repeated recursively until all instances are isolated. This random partitioning produces noticeable shorter paths for anomalies since (a) fewer instances of anomalies result in a smaller number of partitions-shorter paths in a tree structure, and (b) instances with distinguishable attribute-values are more likely to be separated in early partitioning. Hence, when a forest of random trees are collectively producing shorter path lengths for some particular points, then they are highly likely to be anomalies".*



(a)



(b)

FIGURE 15.5 Statistical method: (a) 2D dataset of normally distributed points where X_o is an outlier point (Liu et al., 2012) and (b) interquartile rang (IQR) (Galarnyk, 2020).

Model design: *Sklearn* is used for model design; in addition to the isolation forest algorithm. Both libraries enable the research to encompass the contamination rate; the percentage of an outlier that can be approximately guessed out of total data points. The contaminate rate is determined through the use of the IQR, or interquartile range, as a measure of how widely varying a univariate dataset is. It's the distance between the .25-

quantile and the .75-quantile; also known as "upper" and "lower" quartiles. We consider the middle data segment as normal data points, whereas beyond this range lies outlier points. The IQR method is used on preprocessed economic data (crude oil, DOWIA, S&P 500, gold, and VIX) to collect the contamination rate both in the daily and monthly economic index dataset.

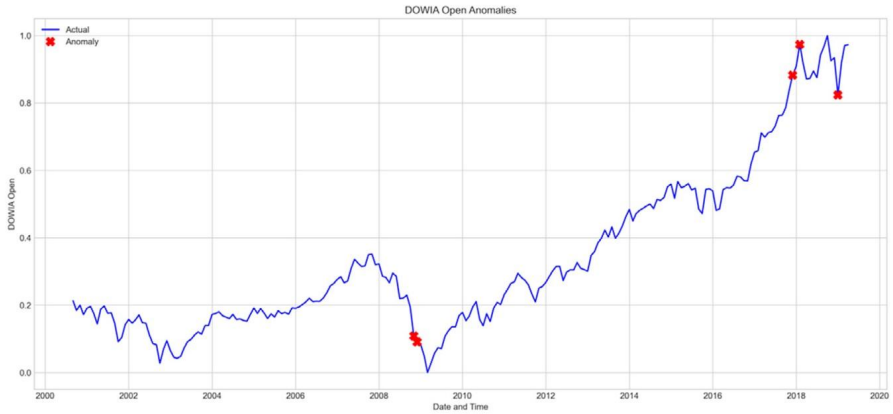
Following the collection of scores and anomalies, red anomalistic points are displayed from the generated isolation forest algorithm. Red anomalistic points are represented by an "x" in Fig. 15.6. The model performance is further tested to reveal a >90% accuracy in terms of labeling the points as the model is supposed to isolate the outliers. See Figs. 15.6 and 15.7.

The graph indicates the anomaly data point distribution from economic indices and major global events with regard to trade/international affairs. Determining the distribution is the act of putting all of the values in a straight line and being able to determine which value has the most density. For instance, the value for VIX open is (0.1), which is approximately the median of the distribution. With one peak, VIX open would be considered univariate, while Gold Open, which contains two peaks instead of one, is multivariate.

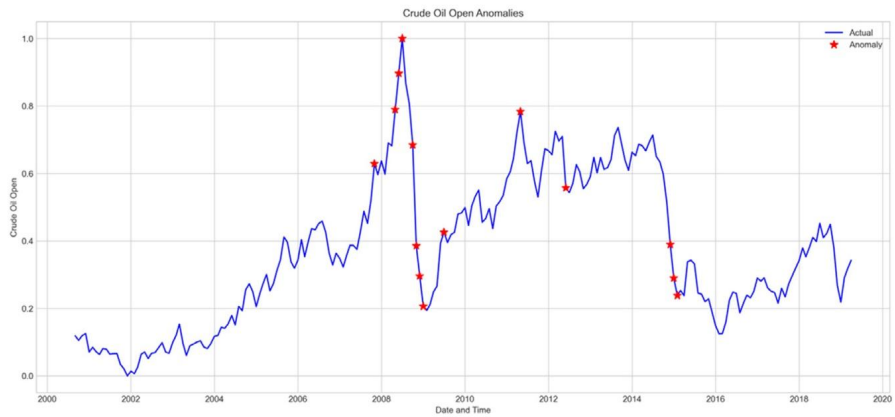
15.2.4 Outlier classification and labeling

Interpreting the red anomalistic points at face value would lead one to believe that those points are the only outlier events within that given year. That observation would be mistaken since the red anomalistic points are only reflective of outliers within that dataset. The purpose of the generated algorithm is to have accuracy in labeling the points as an outlier event, which it accomplished. Ensuring that the outlier events that could not be reflected within the dataset are being captured is the next logical step in this process. To accomplish this, we classify different timeframes as outliers and non-outliers separate from the generated model.

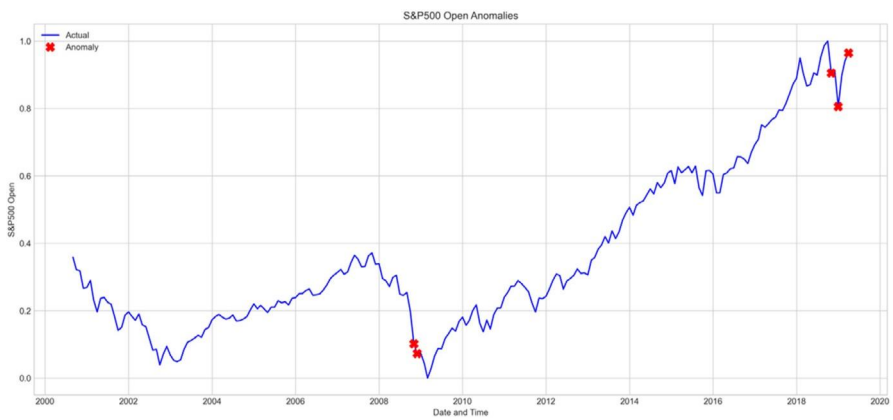
Fig. 15.8 is an illustration of labeled outlier events collected outside of the generated model during the year of 2001. The identification of the outlier events within the figure is not swayed by the time frame per se that the red anomalistic points provide (high peaks and clusters in one area). Rather, the entire year is looked at holistically and all months are considered regard-



(a) DOWIA Open Anomalies

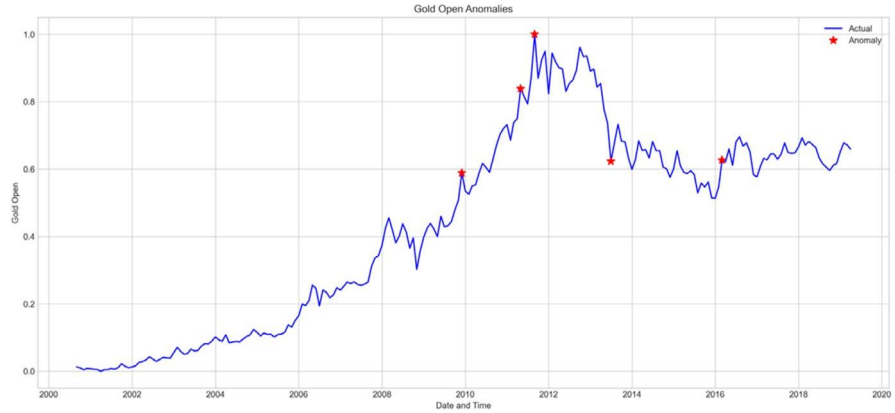


(b) Crude Oil Open Anomalies

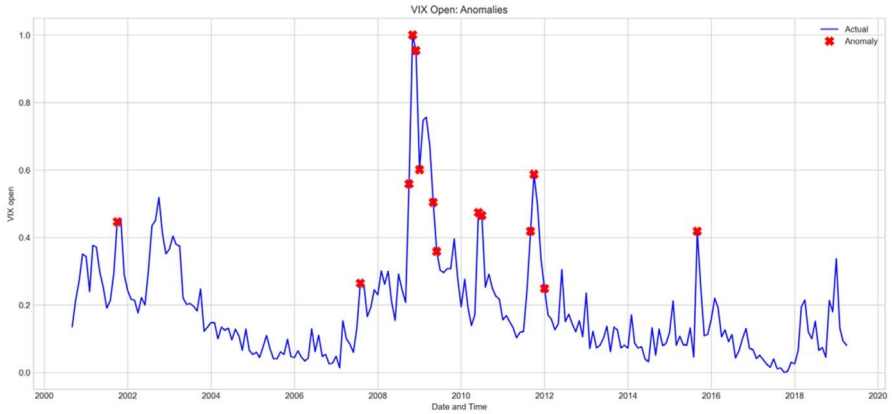


(c) S&P500 Open Anomalies

FIGURE 15.6 Anomalistic data points: (a) DOWIA, (b) crude oil, (c) S&P500, (d) gold, (e) VIX.



(d) Gold Open Anomalies



(e) VIX Open Anomalies

FIGURE 15.6 continued

less if there is a peak of high activity indicated in the generated algorithm beforehand. For instance, the early 2000s recession and 911 attacks are consistent with the generated algorithm, but the trade status shift with China is certainly one that did not fall within the algorithm. The process of going through each year once more to ensure that all possible outlier events are accounted for is repeated for each year.

Afterwards, the outlier events are filtered and classified into one of the following categorization ID's (Financial=1, Global=2, Pandemic=3, Political =4, Weather=5). The act of classifying an outlier event broadens the narrow scope of weather that agriculture is accustomed to. When this infor-

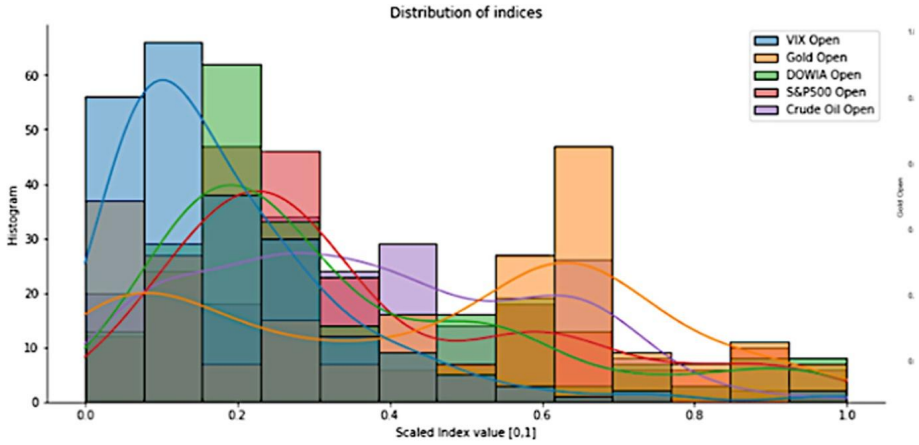


FIGURE 15.7 Distribution of indices.

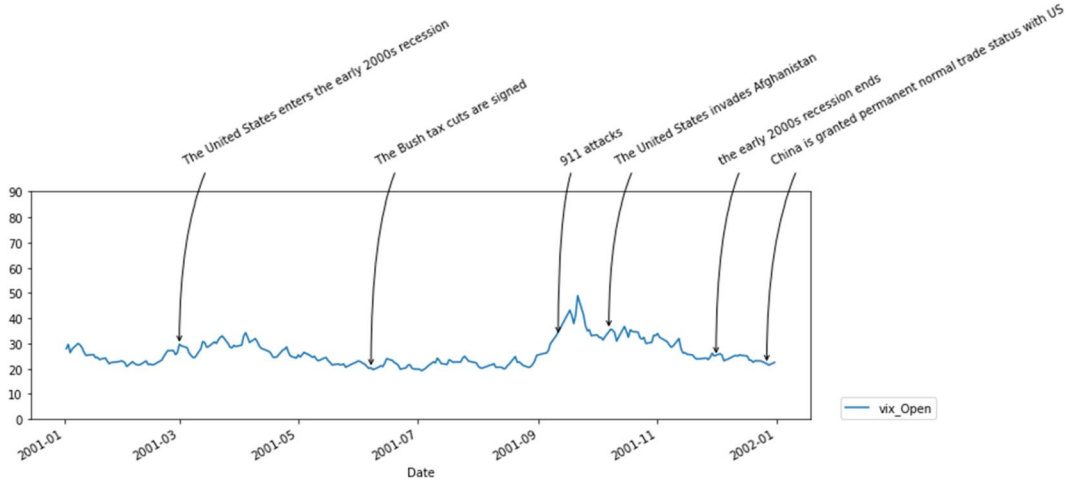


FIGURE 15.8 Labeled outlier events.

mation coincides with fluctuation input from economic indices, it places farmers in a better position to make better decisions. In that sense, the AI2Farm model appeals to the liking of Verdouw et al. (2015) who depicted the following management functions:

1) analysis and decision-making: comparing measurements with the norms that specify the desired performance (system objectives concerning e.g.,

quantity, quality, and lead time aspects), signaling deviations, and deciding on the appropriate intervention to remove the signaled disturbances.

2) intervention: planning and implementing the chosen intervention to correct the farm processes' performance.

15.3 Insight into data driven farming

The intent of each site visit is to develop a better understanding of factors that are a hindrance to yield production and efficiency on farms across the Virginia Tech network as a whole. Variance in opinion occurred due to each farm specializing in a different aspect of agriculture from one another. Kentland Farm focused on (crops, breeding, plants), McCormick Farm focused on (cattle, feed, Animal Science), and the dairy complex focused on (Dairy Science). While in attendance, we observe what technology is already in use as well as what technology could be put in place to improve upon the current conditions. We then generalize responses as qualitative data to complement the preexisting categorization ID's (Financial=1, Global=2, Pandemic=3, Political =4, Weather=5).

For instance, a small grain breeder at Virginia Tech shared an example of a severe weather event example during the visit. In Virginia, wheat and barley are planted in the fall, go dormant over the winter, and come back in the spring to produce grain. When maturity happens, weather has an effect on whether or not it can be pulled out of the field in time or if they have the quality that's necessary to be a viable crop. Once grains are mature, barley specifically, the grain dries down and becomes ready to become a seed, carrying a certain level of dormancy with it. However, if they are rained on within the field, then they'll rehydrate and sprout within the field, while they're still in the grain head, which reduces the quality of the product. This severe weather event example added validity to the need of supporting smart farm initiatives to integrate new technologies into farming practices.

15.3.1 Kentland and dairy farm

Crop Management through the use of AI is actively being used at Kentland Farm. The method of crop management entailed; pre-mapping of land and crops, drone calibration, and navigation using GPS (Global Positioning Sys-

FIGURE 15.9 Prepared farm via vertical and horizontal lines for drone calibration.

tem). The quality of soil is ever so important for managing crop yields (Ge et al., 2011). To have an accurate depiction of the quality of the soil within the field, the farm is separated into equal boxes. Vertical and horizontal lines are then constructed for drones to be able to detect areas within the farm to know where the end and the beginning is and to provide data in that specific area. The farm is then prepared for the drone to fly above it and be able to take images and read every piece of land separately.

Figs. 15.9–15.13 are sample images from Smart Farms at Virginia Tech.

15.3.2 Shenandoah Valley Agricultural Research and Extension Center (SVAREC)

The Shenandoah Valley Agricultural Research and Extension Center (SVAREC) conducts pasture system research and beef cattle production within the confines of over 900 acres of owned and leased land. Cattle are used for breeding and various projects based upon (artificial insemination, weight, body condition scores, hip height, pregnancy checks, sex of the fetus, weight of the calf, etc.).

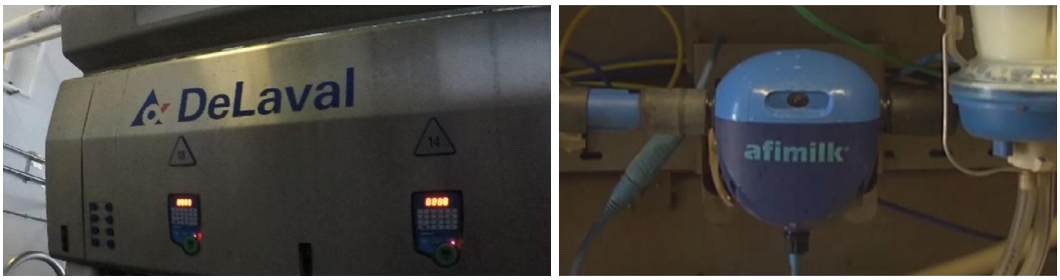
FIGURE 15.10 Wide-angle view of weather stations at SVAREC.

(a) SmartScale (b) Mechanical Feeder

FIGURE 15.11 (a) SmartScale and (b) mechanical feeder at McCormick farms.

FIGURE 15.12 Interconnected outlier event.

The large overlay of the farm includes the weather station at the top right corner as its focal point. Weather stations are a common piece of technology on farms used primarily for measurements of precipitation, air



(a) DeLaval Machines (b) Afimilk Device

FIGURE 15.13 Dairy machinery installed at the farm: (a) DeLaval and (b) Afimilk.

temperature, dew point temperature, wind speed & direction, barometric pressure, soil temperature & moisture, and solar radiation. Additionally, virtual fencing accompanies this large area of land. Virtual Fencing "contains cattle by providing audio and electrical signals via a neckband device and assists in measuring activity, health variables, weight, location, movement towards water, and feed management. Animals are restricted in a specified area via receiving stimulatory cues, rather than through the presence of a physical fence enabling remote animal monitoring and movement control" (Keshavarzi et al., 2020). Enhanced mobility of cattle is an area in which farmers at Kentland are keen to continue to expand upon, which will be a part of future research.

Cattle control, having agency over cattle to produce a desired outcome or act in a desirable manner, is currently functioning as intended due to the following two technologies in place. The mechanical feeder is in use outside of the traditional sense during the pandemic to slow feed for slower slaughtering dips in consumption or national demand (top left) and the SmartScale is used for weight management. Data collected from such measures include the following: body condition, hip height, age, calving dates, hay amounts, feed costs, and weather data. All of which is part of economic analysis for public policy toward the Farm Bill (USDA, 2021). The definition of SmartScale and virtual fencing is as follows (Producer Smart Scale, year):

> *"SmartScale is a wireless scale system that captures animal weight, performance, and behavior each time it drinks water. SmartScale is a cloud*

connected, automated scale unit that utilizes existing pen water supplies to provide daily weight and performance for each animal in your pen. Customizable to fit most existing pen water supplies and integrates with SmartFeed bunks to provide high-quality, real-time data".

Supply chain bottleneck is an example once more of how there's an overlapping of outlier events. Supply chain bottleneck is defined as congestion in the production system due to an increase in demand with limited capacity. The result, in this case, is supply overstock of cattle at a weight prepared for slaughter at a state too early. Supply chain bottleneck occurred when some meat processing plants shut down, preventing animals awaiting slaughter from being processed. Cattle producers had a "12.3% decrease in the price they receive. Although producer and consumer prices tend to move in unison, the supply-chain bottleneck caused by Covid-19 has likely caused a divergence" (Beckman and Countryman, 2021). Hence, why there was and still is a need for the feed management tool.

With this scenario, one will find that outlier events are overlapping in four out of the five categories. The pandemic impacted the world, which in turn negatively impacted the market, which in turn negatively impacted the farmers in such a way that involvement of the US Department of Justice (DOJ) was needed.

Although the DOJ has reportedly contacted the four big meatpackers (Tyson Foods, JBS SA, Cargill, and National Beef) to seek information related to an investigation into possible antitrust violations, concerns of price fixing during the pandemic will continue to mount (Johnston, 2020). As long as there is an imbalance and presence of middlemen in between farmers and consumers, profits will never make it down to small farmers, resulting in farmers across the country continuing to not get their fair share. The notion of living amidst a "broken market" due to anti-competitive practices and market manipulation by the meat packing industry rings true. This multifaceted outlier event will continue to impact farmers and constrain farmers to use technology such as the food monitor to alleviate the supply chain bottleneck when in fact the financial aspect of the diagram above is the root cause of the problem.

Now, that's not to say that there aren't outlier events that stand on their own. The identification of such events, especially outside of the typical weather outlier event, is equally as important. The ability to understand outlier events and to view them in this manner will aid in the decision-making process for farmers during both conventional and unconventional times. Thus making the AI2Farm model even more justifiable for its use in the future.

15.3.3 Dairy complex at Virginia Tech's SmartFarms

The dairy complex processes 2k gallons of milk every two days through automated milking. The data collected includes herd analysis, daily data for production, cow's health, activity tracking, milk quality, and infections. DeLaval machinery analyzes milk samples from lactating dairy cattle through somatic cell count (SCC) to monitor udder health and diagnose subclinical intramammary infection (IMI) in dairy cattle (Kandeel et al., 2017). The AfiMilk system is versatile and can be of use for their ICAR milk meter, integrated farm management SW, heat detection system, and milk analyzer (Berger and Hovav, 2013).

The overall consensus is that the newer technology is solely being utilized on the smart farms due to its function as a test bed for the development and testing of technologies; in other words, it's slower to be adopted by beef cattle producers outside of the Virginia Tech network. The hesitancy displayed by other farmers is due to a lack of trust. Trust in precision agriculture is dependent on recommendations that are "reliable, accurate, transparent, and fair than previous systems" (Gardezi and Stock, 2021). Essentially, requesting farmers to place trust in an algorithm or model which they are not familiar with and goes against traditional modes of farming that have been established over the years is bold from the researcher's standpoint. Farmers are no longer the sole reserve of experience, as cognition and decisions have increasingly become distributed between farmers and intelligent technologies.

15.4 Larger policy implications

Public policy is a course of government action or inaction taken in response to social problems. When written down as laws and directives, it serves as precedent in future cases to see whether or not the policy was being upheld or not. The government is a collective since it is not a unilateral decision-making entity (made by one person) and, under a democratic form of government, is *consensual* since citizens can elect who has the privilege of making decisions on their behalf.

The process of public policy begins with agenda setting (prioritization); how issues get defined as political and worthy of government attention or action by elected officials. Next is policy formulation, when a piece of legislation can be introduced as a bill by a certain congressman, through signature of an executive order by the president, or by bureaucratic entities. Following this step is policy implementation, the time lapse effects of such policy being in place. Lastly, policy evaluation uses data to see if the policy is working as intended.

In the political sphere, efforts to address agricultural issues have been made in an overarching manner. In *The fault lines of farm policy*, Coppess (2018) contended that farm risk is made up of two fundamental matters: market risks (whether lost export demand or oversupplied markets. These risks return prices too low to cover cost and profit needs) and weather risks (the dilemma between good weather that can result in massive crops that outstrip demand and lower prices and bad weather that can cause damage to crops that leave too little to cover costs and needs) (Coppess, 2018). Individuals and organizations who advocate on behalf of these concerns include the Secretary of Agriculture, the USDA, agriculture committees within Congress, farm lobbyists, and others. Concerns may be vocalized as a means in which to combat activities perceived as detrimental or an endangerment to society, to protect certain populations and or groups within society, or to promote certain activities that are deemed important.

Such concerns are expressed over the years and culminate into what is known as the *Farm Bill*. United States agricultural policy generally follows a 5-year legislative cycle producing a *Farm Bill* with the Agricultural Improvement Act of 2018 (Congressional Research Service, 2019) being the most

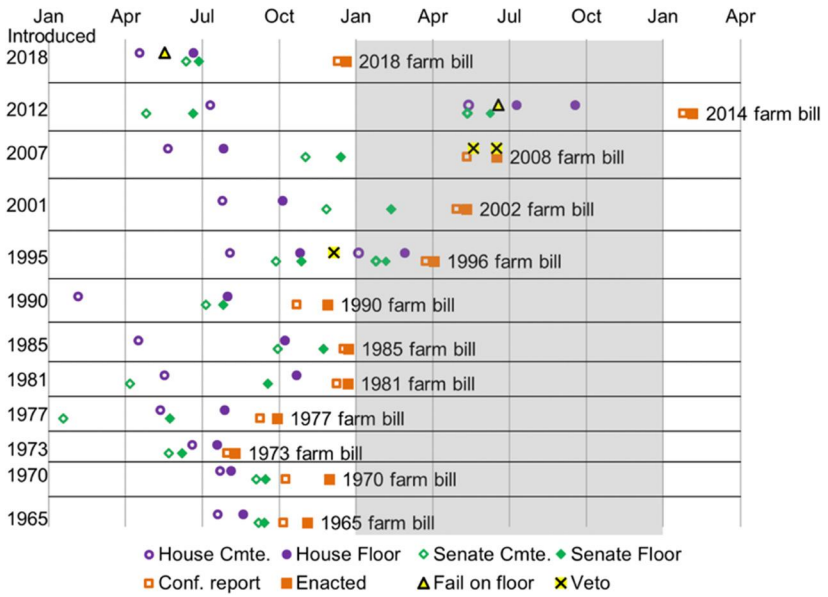


FIGURE 15.14 Major legislative actions on farm bills, 2018–1965 (Congressional Research Service, 2018b).

recent. Historically, there has been a trend of enactment occurring well after the original expiration dates. The possible consequences of expiration include "minimal disruption (if the program is able to be continued via appropriations), ceasing new activity (if its authorization to use mandatory funding expires), or reverting to permanent laws enacted decades ago (for the farm commodity programs)" (Congressional Research Service, 2018a). What's more, is that Farm Bill reauthorization has become more complex with the process of enacting a new farm bill varying from previous years as follows (Congressional Research Service, 2018b) (see Fig. 15.14):

> *"Prior to the expiration of the existing law has become more difficult. As stakeholders in the farm bill have become more diverse, more people are affected by the legislative uncertainty around this process. This lack of certainty may translate into questions about the availability of future program benefits, some of which may affect agricultural production decisions or market uncertainty for agricultural commodities."*

In regard to outlier events, the Farm Bill contains support programs for agricultural disaster assistance, such as Price loss coverage (PLC), Agriculture risk coverage (ARC), and the Marketing assistance loan (MAL) program. However, federal assistance to recover financially from natural disasters is a method that occurs after the fact and is not a preventative measure desired by most farmers. Responding in the manner of federal assistance qualifies as a short-run policy (primarily in the coming weeks) over which the "supply of goods and services can be altered into a better state for essential goods and services" (The Brookings Institution and Snower, 2020).

Specific efforts to accommodate farmers' needs during the pandemic included funding such as the American rescue plan, the U.S. Department of Agriculture's implementation of the Coronavirus Aid, Relief, and Economic Security (CARES) Act of 2020, and much more (USDA, 2021). The pandemic was chosen in large part due to its illustration of how intricate and interconnected the market is. The global ramifications of the spread of the coronavirus were evident when policy adjustments were made simultaneously across the globe; adversely disrupting market and trade. An analysis of the distribution of measures undertaken by 54 countries during the first four months of 2020 provides some early insights into the emphasis, scope, and regional diversity of policy responses. The study found that temporary measures taken by existing countries within the international relations community, had "adverse effects on consumers (import restrictions or local promotion measures), producers (export restrictions), food chain actors (market distorting measures), and the environment (regulatory relaxations, input subsidies)" (Gruère and Brooks, 2021). In turn, temporary relief measures as a response to outlier events, is a double edge sword. Lifting measures at the conclusion of such an event not only will send the market into shock once more, but complicate the relationship of actors in the future.

Presenting information from the AI2Farm model to policymakers (U.S. government) would alleviate the need to disperse funding affecting the national budget on such short notice and lessen the reliance upon loans for commodities by farmers. Additionally, the data on outlier events could be used as evidence within cases such as the disparity between packers' prof-

its and beef prices which have widened during the pandemic being brought before the DOJ, quickening the process for policy changes that would most likely occur before the typical 5-year legislative process.

The Farm Bill is faced with the daunting task of not only improving upon the bill from previous years, but also navigating uncertainty with the implementation of new procedures in the future. Despite the best effort of a well written bill, the inevitable zone of uncertainty diminishes the impact of such policies (Novak et al., 2015). Therefore successful production season would be simply unattainable if farmers weren't afforded the flexibility provided by the AI2Farm model and had to rely heavily on the details stated within the Farm Bill.

15.5 Conclusion

In this work, we posed the following question that is inverse of the typical way that agriculture farming is discussed: One should ask not what is the most efficient way to provide aid to farmers in the form of compensation for commodity and income losses following an outlier event, but rather what is the most efficient way to inform farmers about conventional and unconventional time to alleviate shock to commodity production. The AI2Farm model provides farmers with the much needed flexibility to persist within the ever changing environment within society.

Acknowledgments

We would like to acknowledge the SmartFarm Innovation Network for catering to our research needs throughout the investigative and writing processes, including members from the Center for Advanced Innovation in Agriculture (CAIA) at Virginia Tech. We would like to thank The Commonwealth Cyber Initiative (CCI) at Virginia Tech for providing opportunities for students from the regional institutions to conduct AI research.

References

Becker-Reshef, I., Justice, C., Sullivan, M., Vermote, E., Tucker, C., Anyamba, A., et al., 2010. Monitoring global croplands with coarse resolution Earth observations: the global agriculture monitoring (GLAM) project. Remote Sensing 2 (6), 1589–1609. https://doi.org/10.3390/rs2061589.

Beckman, J., Countryman, A.M., 2021. The importance of agriculture in the economy: impacts from COVID-19. American Journal of Agricultural Economics 103 (5), 1595–1611. https://doi.org/10.1111/ajae.12212.

Berger, R., Hovav, A., 2013. Using a dairy management information system to facilitate precision agriculture: the case of the AfiMilk® system. Information Systems Management 30 (1), 21–34. https://doi.org/10.1080/10580530.2013.739885.

Brown, C.M., Vostok, J., Johnson, H., Burns, M., Gharpure, R., Sami, S., et al., 2021. Outbreak of SARS-CoV-2 Infections, Including COVID-19 Vaccine Breakthrough Infections, Associated with Large Public Gatherings — Barnstable County, Massachusetts, July 2021. Morbidity and Mortality Weekly Report 70 (31), 1059–1062. https://doi.org/10.15585/mmwr.mm7031e2.

Congressional Research Service, 2018a. Expiration of the 2014 farm bill. https://sgp.fas.org/crs/misc/R45341.pdf.

Congressional Research Service, 2018b. Farm bills: major legislative actions, 1965–2018. https://sgp.fas.org/crs/misc/R45210.pdf.

Congressional Research Service, 2019. The 2018 farm bill (P.L. 115–334): summary and side-by-side comparison. https://crsreports.congress.gov/product/pdf/R/R45525.

Coppess, J., 2018. The Fault Lines of Farm Policy: A Legislative and Political History of the Farm Bill. University of Nebraska Press, Lincoln, Nebraska.

D'Agostino, A.L., Schlenker, W., 2016. Recent weather fluctuations and agricultural yields: implications for climate change. Agricultural Economics 47 (S1), 159–171. https://doi.org/10.1111/agec.12315.

Elleby, C., Domínguez, I.P., Adenauer, M., Genovese, G., 2020. Impacts of the COVID-19 pandemic on the global agricultural markets. Environmental & Resource Economics 76 (4), 1067–1079. https://doi.org/10.1007/s10640-020-00473-6.

Galarnyk, M., 2020. Understanding boxplots - towards data science. Retrieved from https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51.

Gardezi, M., Stock, R., 2021. Growing algorithmic governmentality: interrogating the social construction of trust in precision agriculture. Journal of Rural Studies 84, 1–11. https://doi.org/10.1016/j.jrurstud.2021.03.004.

Ge, Y., Thomasson, J.A., Sui, R., 2011. Remote sensing of soil properties in precision agriculture: a review. Frontiers of Earth Science. https://doi.org/10.1007/s11707-011-0175-0.

Gruère, G., Brooks, J., 2021. Viewpoint: characterising early agricultural and food policy responses to the outbreak of COVID-19. Food Policy 100, 102017. https://doi.org/10.1016/j.foodpol.2020.102017.

Johnston, M., 2020. Timeline. Fair Cattle Markets. https://fair-cattle-markets.com/timeline/.

Kandeel, S., Megahed, A., Arnaout, F., Constable, P., 2017. Evaluation and comparison of 2 on-farm tests for estimating somatic cell count in quarter milk samples from lactating dairy cattle. Journal of Veterinary Internal Medicine 32 (1), 506–515. https://doi.org/10.1111/jvim.14888.

Keshavarzi, H., Lee, C., Lea, J.M., Campbell, D.L.M., 2020. Virtual fence responses are socially facilitated in beef cattle. Frontiers in Veterinary Science, 7. https://doi.org/10.3389/fvets.2020.543158.

Knoema, 2021. Cost of production for major field crops in U.S. - knoema.com. Retrieved from https://knoema.com/USDACPFC2021Jun/cost-of-production-for-major-field-crops-in-u-s.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data 6 (1), 1–39. https://doi.org/10.1145/2133360.2133363.

Loukas, S., 2021. Everything you need to know about min-max normalization: a Python tutorial. Medium. https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79.

Novak, J., Pease, J., Sanders, L., 2015. Agricultural Policy in the United States: Evolution and Economics, 1st ed. Routledge.

Purdue University Agricultural Economics, 2020. FoodandAgVulnerabilityIndex. Retrieved from https://ag.purdue.edu/agecon/Pages/FoodandAgVulnerabilityIndex.aspx?fbclid=IwAR2YKurZ3Q0PSWMI8b3aotyD9FvzKXqRYBOQyMTmpiDbXiviGVVVutBME10.

Shaikh, I., 2019. On the relationship between economic policy uncertainty and the implied volatility index. Sustainability 11 (6), 1628. https://doi.org/10.3390/su11061628.

Sharma, A., Kiciman, E., 2020. DoWhy: an end-to-end library for causal inference. arXiv preprint. arXiv:2011.04216.

The Brookings Institution, Snower, D., 2020. The socioeconomics of pandemics policy. https://www.brookings.edu/wp-content/uploads/2020/04/socioeconomics_of_pandemics_policy.pdf.

United States Department of Agriculture, 2018. Federal risk management tools for agricultural producers: an overview. https://www.ers.usda.gov/webdocs/publications/89202/err-250.pdf?v=0.

USDA, 2021. In historic move, USDA to begin loan payments to socially disadvantaged borrowers under American rescue plan act Section 1005. https://www.usda.gov/media/press-releases/2021/05/21/historic-move-usda-begin-loan-payments-socially-disadvantaged.

Vandello, J.A., Cohen, D., 1999. Patterns of individualism and collectivism across the United States. Journal of Personality and Social Psychology 77 (2), 279–292. https://doi.org/10.1037/0022-3514.77.2.279.

Verdouw, C., Beulens, A., Reijers, H., van der Vorst, J., 2015. A control model for object virtualization in supply chain management. Computers in Industry 68, 116–131. https://doi.org/10.1016/j.compind.2014.12.011.

This page intentionally left blank

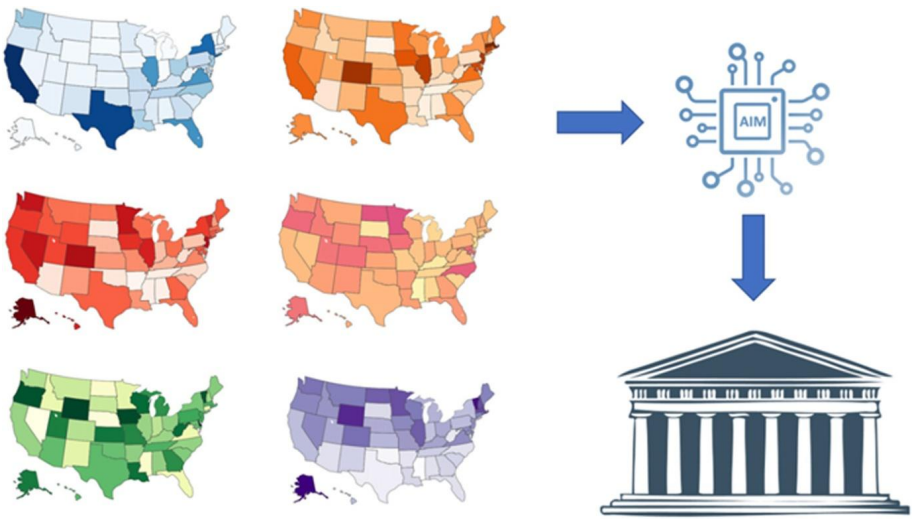
16

Bringing dark data to light with AI for evidence-based policymaking

Dominick J. Perini[a], Feras A. Batarseh[b], Amanda Tolman[c], Ashita Anuga[d], and Minh Nguyen[e]

[a]Northrop Grumman, Denver, CO, United States [b]Department of Biological Systems Engineering (BSE), College of Engineering (COE) & College of Agriculture and Life Sciences (CALS), Virginia Tech, Arlington, VA, United States [c]Radford University, Radford, VA, United States [d]Commonwealth Cyber Initiative, Virginia Polytechnic Institute and State University, Arlington, VA, United States [e]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States

Graphical abstract



Abstract

Addressing the problem of "Dark Data" is a challenge that is faced in all industries; this chapter proposes a solution designed for the policymaker. The legisla-

AI Assurance. https://doi.org/10.1016/B978-0-32-391919-7.00030-5

tive process has long been fueled by countless hours of research into the past as well as predictions of the future. The effort of manual information processing is positioned to be re-engineered for the information age due to progress made in big data and Artificial Intelligence (AI). The system presented in this chapter designated as the "Associated Impact Measure," or AIM, integrates spatiotemporal data collected from state repositories with technology-related public policy information into a single dataset. The policies are then compared textually using natural language processing (NLP) methods and numerically with the use of environmental descriptors; data are collected on the states themselves. Finally, these comparisons are utilized by an artificial neural network to learn and predict the associated impacts that the policies have on trends describing technology usage for the respective state. The primary struggle of creating legislation is determining how it will affect the state/country. Addressing dark data as a means of AI assurance, as well as considering the ethics of using AI in the legislative process are of the utmost concern; in this chapter, they are measured, experimented, and presented.

Keywords

AI applications, dark data, evidence-based policymaking, artificial neural networks, natural language processing, spectral clustering

Highlights

- Using TFIDF and spectral clustering, legislation text is represented numerically, and similar laws are grouped into meaningful clusters
- An artificial neural network is used to make nonlinear inferences on an atypical dataset with high accuracy
- Dark data surrounding the policymaking process is incorporated into an AI-enabled system that produces more accurate predictions of trends in technology usage
- Evidence-based policymaking is benefited by AI-enabled systems that take into consideration contextual and multidimensional data that are linked together spatiotemporally
- Ethical considerations in certain AI applications are necessary, especially when the domain of application is as human-centric as public policy
- This chapter concludes with a discussion about AI's place in the past, present, and future of the legislative process

16.1 Introduction

The legislative process does not operate in a vacuum, but instead at the intersection of many domains (Klein, 1990). This chapter presents an explanation and demonstration of a method of predictive modeling that uses cutting-edge practices in many fields, such as engineering, political science, mathematics, and social sciences. When examining the scope of public policy, the many phases that the laws at hand go through are research, writing, negotiation, voting, enactment, and enforcement; see Keating (2016). Throughout this process, existing laws are used to inform the needs of future laws by identifying gaps or overreaches that they might contain. Measuring and quantifying the impact that a policy has on the world is a difficult and persistent challenge (Segone, 2008). It is this exercise that we propose can be assisted and improved using AI. With these primary considerations, a clear need can be identified for a system that facilitates technological advancements in policy advocacy.

With the goal of developing an appropriate solution to the aforementioned problem, data are collected and we develop a pipeline that aggregates the text of public policies related to technology and data collected from several national sources into one dataset. Through the lens of an ANN, the dataset is used to support evidence-based policymaking for legislators by way of providing them with AI-derived insights about the policies they are interested in.

16.1.1 Background

When establishing a framework for merging data and legislation, several concerns need to be addressed. AIM is developed in a way that includes ever-present "dark data" in the legislative research process by expanding the scope from which information is gathered, which is one of the primary considerations when addressing the reliability and accuracy of AI systems. Dark data as a class of information is not readily accessible or stored in a way that makes it practically invisible to scientists and other potential users, therefore the information is likely to remain underutilized and eventually lost (Heidorn, 2008). As the promise of value from big data grows, and cloud storage becomes less expensive and more accessible, data are being

hoarded and stored for every possible topic or subject. The existence of this information is not causing problems, but the conclusions that intelligent systems come to without the "dark data" might be, for more information refer to Grimm (2018).

This is one of the most important concepts related to AIM, because the purpose of the system is to merge readily available data about technology usage with public policies and make predictions using that information, which is directly addressing dark data. To take advantage of the data that are collected, the developed framework for the project is designed to append contextual information about when and where the law was passed. Including this additional data as inputs to the model makes for a well-informed and reliable method of predictive modeling.

The overarching goal of the AIM system is to automate and perform policy advocation. The term "policy advocacy" is used to describe any research or process that terminates in the direct advocacy of a single policy or a group of policies (Gordon et al., 1977). It would be possible for human researchers to go through the data available online and identify trends, make associations, and come to conclusions about future policies; however, this is a cumbersome and delicate process. The benefits of using AI for this type of application are that new patterns can be recognized that might not be intuitive or obvious to a human observer, and the computations can be done in seconds, rather than days or weeks of research by humans. AIM exists at the intersection of policy monitoring and big data, and should be used as a tool for political researchers when performing policy evaluation.

16.1.2 Motivation

In the wake of a global pandemic, policymaking has been scrutinized and public trust in opinion-based policymaking has been diminished (O'Mathúna et al., 2021; Yaros et al., 2021). It is more important than ever to begin integrating intelligent systems into the policy evaluation and advocacy process. Evidence-based approaches facilitate policymaking in a manner that is more respective of the scientific method (Nutley et al., 2002), beginning with initial hypotheses and evaluating them based on experimentation and research. On the contrary, opinion-based policymaking is typ-

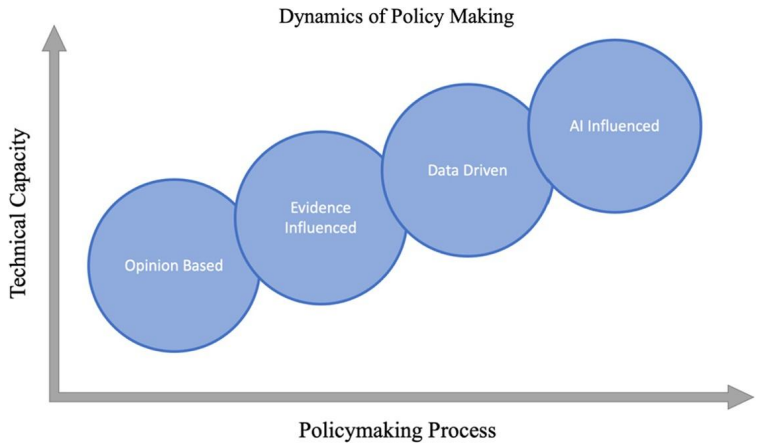


FIGURE 16.1 Dynamics of Policy Making: There is overlap between each type of political process, with each succeeding process having increased technical influence. Adapted from Segone (2008).

ically based on limited and selective use of evidence or untested views of individuals or groups, based on ideology, prejudice, or speculative conjecture. Fig. 16.1 illustrates that there is a spectrum of policymaking dynamics, and that evidence-based politics is only possible with an increased technical capacity, which AIM is positioned to support. The original figure, see Segone (2008), has three dynamics: opinion-based, evidence-influenced, and evidence-based. We see fit that evidence-based dynamics can be broken into two distinct policy dynamics, data-driven and AI influenced.

Each policy-making dynamic has a technical capacity that is low in totally nontechnical processes, and high in very technical processes. As the technical capacity of a dynamic increases, so does the amount of abstraction between evidence and policy decisions. Opinion-based can be thought of as no level of abstraction between a lawmaker's experiences and opinions, and their policy decision. An evidence-influenced process incorporates some amount of evidence outside personal experience in the logic of the process. The data-driven process expands the "evidence" further to include historical trends, data visualizations, and statistical analysis of metrics to inform a policy decision. Where AIM falls on this spectrum is under the AI influenced category, because there is a middle-man between the data and the policymaker, providing an additional level of abstraction necessary

for such complex and large datasets, and that middle-man is the AI algorithm itself.

AI research is a broad term that can mean several different things, especially when comparing its usage across industries (Batarseh et al., 2021). The breadth of applications of AI is constantly expanding, and AI for public policy is on the frontier of that expansion. Many companies and institutions involved with AI are putting in place frameworks and guidelines that can direct these new developments in safe and human-centric ways (Thierer et al., 2019). One of the most prominent groups dedicated to exploring the use of AI systems in policymaking processes is the Alan Turing Institute, where their Public Policy research program states their mission as "Working with policymakers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making" (Leslie et al., 2021). It is initiatives like this that inspire and encourage the research done in this chapter, because it becomes clear that there is an interest and demand for high-quality AI-powered systems that can be used to elevate the policymaking process for the good of all.

16.1.3 The AIM pipeline

As a prerequisite to discussing any individual component of the AIM system, it is important to understand the overall flow of information and how it is transformed and combined. The pipeline adheres to a traditional supervised machine learning structure, using features and labels to train a model, and then predicts the labels based on unseen features. In this case, the input features are a combination of two types of data: individual policies and environmental descriptors. Environmental descriptors numeric features describe the "environment" that the law was passed in; this will be explained further in Section 16.2.4. These features are then married to labels, which are metrics of interest that are tracked over space and time (state and year). Each datapoint's feature set describes the text of the law as well as the circumstances of its enactment, each set of labels represents an associated impact on a given metric of interest; this will be explained further in Section 16.2.2. The neural network is trained on this data and predicts those labels for new policies, and these predictions are used to better

understand the effect that upcoming or proposed policies will have on the world around them. The white boxes in the diagram represent transformations in the data, which are necessary when combining data that come from different sources and are different data types.

16.2 The dataset for AIM

The foundation of any AI system is a dataset architecture that allows for inference and understanding, and that was a primary consideration for the development of AIM. Data are collected from many different sources, ranging from state websites to government agencies. Having a diverse pool of information poses both challenges and opportunities that are based on the quality of the connection between data. One of the primary concerns that arose from the data collection process was answering the following question: How will the data be connected? To find out, a paradigm had to be developed that could join together data describing technology usage, legislation, and state information.

16.2.1 Dataset paradigm

The AIM system starts with unstructured, unlabeled data. Three types of data that come from numerous sources are hard to link among a common dimension, especially one that makes sense in the context of the problem. Using AI for policy recommendation, this connection needs to be a significant one. For this reason, we decided to connect the data spatiotemporally. This common thread is visualized in Fig. 16.3, showing the three categories of data and examples of their respective dimensions. Technology usage metrics and state information are tracked regularly over time, and each state is tracked separately. Each individual policy includes the corresponding state and the year it was passed in. Therefore the commonality is that these data are connected in space and time, being year and state. This allows the specificity of each metric value to be assigned to specific laws, for example, a law passed in 2017 for the state of Virginia will have environmental descriptor and technology metric values specific to that state and year.

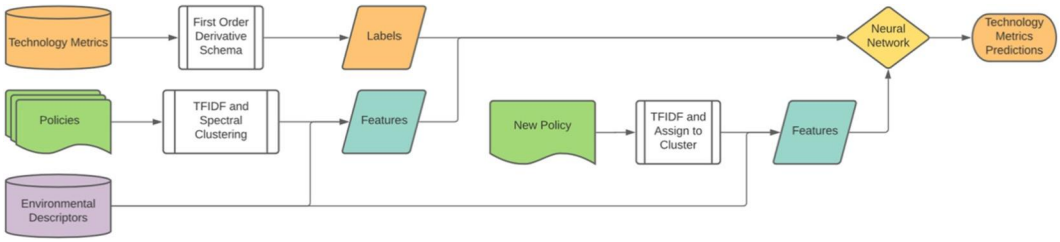


FIGURE 16.2 AIM pipeline diagram – Colors represent data types.

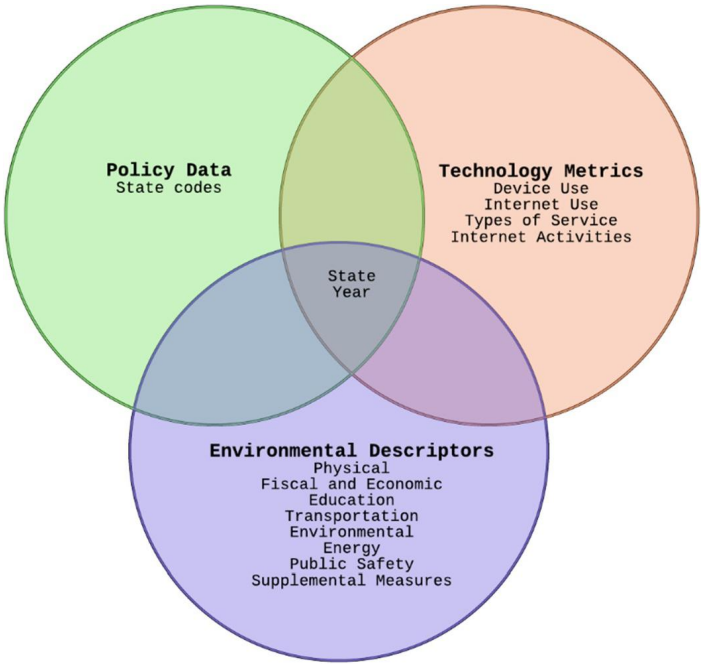


FIGURE 16.3 Technology Policy Data Venn Diagram: The common feature in the data is that all variables are associated with a state and a year.

In Fig. 16.2 and Fig. 16.3, data categories are represented graphically. Orange-colored (mid gray in print version) sections are metrics of interest or the labels for the data and are used to assess the impacts laws had on the world; green-colored (light gray in print version) sections are legislation texts and are how laws are represented; blue-colored (gray in print version) sections are features for the data, which are engineered to be used as inputs to an algorithm, and the purple-colored (dark gray in print version) section

is environmental descriptors, which are additional inputs that address dark data.

16.2.2 Metrics of interest

Metrics of interest are data that describe how people use technology across the United States. The purpose of using this information is to understand trends in technology usage and how things change over time, and how people use technology differently in different parts of the country. Table 16.1 provides an exhaustive list of all of the metrics of interest used for training and prediction with AIM. One example might be that people in Maryland might use the internet to work remotely more than people who live in Washington, DC. By predicting these metrics for upcoming policies, policymakers can be informed on the potential impact their law might be associated with for key metrics that they deem important to their constituent's interest.

As a means of determining policy efficacy, metrics of interest can be used on their own or in conjunction with AIM to make future predictions about how metrics of interest might be impacted by a piece of legislation. For example, a state legislator in Vermont might be interested in passing a bill that would increase the proportion of citizens that have access to broadband internet, and AIM's analysis would be able to tell that legislator that the law they give as an input is predicted to increase that metric by 50,000; this would be a positive indicator for the policy's potential. Each metric has a minimum temporal resolution of 1 year and is available for all 50 states and Washington, DC.

In total *47 metrics* of interest are all tracked over many years and for each of the 50 states. The overall theme of the metrics is that they can reveal patterns in the ways that people use technology, which is what we suspect the laws will have an influence on, because all of the laws are related to technology. If a change in the metric spatiotemporally coincides with the passing of a policy, and this happens consistently with similar policies, then it would be an important insight to lawmakers and is the bedrock of this project.

Table 16.1 List of Tracked Metrics of Interest: Sourced from US federal census survey responses.

Metrics of Interest
Device Use: Universe: Ages 3+ Civilian Persons
• The Number of Civilians That Use a Desktop Computer • The Number of Civilians That Use a Laptop • The Number of Civilians That Use a Tablet or An E-Book • The Number of Civilians That Use a Smart Tv or A TV - Connected Device • The Number of Civilians That Wear Smart Watches, etc
Internet Use: Universe: Ages 3+ Civilian Persons
• The Number of Internet Users That Are Ages 3 And Up • The Number of Adult Internet Users That Are Ages 15 And Up • The Number of Civilians Who Use the Internet at Home • The Number of Civilians Who Use the Internet at Work • The Number of Civilians Who Use the Internet at School • The Number of Civilians Who Use Internet at Coffee Shop or Other Business • The Number of Civilians Who Use the Internet While Traveling Between Places • The Number of Civilians Who Use the Internet While in A Public Place (Library, Community Center, etc) • The Number of Households That Has Anyone in The House to Use the Internet at Any Location • The Number of Households That Use Home Internet While at Home • The Number of Households That Have No Home Internet Use by Anyone Member in The House
Non-Use of the Internet at Home: Universe: Households Without Any Home Internet Users
• The Number of Offline Households That Had Prior Home Internet Use by Anyone in House • The Number of Offline Households Whose Main Reason for Offline Is No Need or Interest • The Number of Offline Households Whose Main Reason for Offline Is Internet Is Too Expensive • The Number of Offline Households Whose Main Reason for Being Offline Is No Computer • The Number of Offline Households Whose Main Reason for Being Offline Is They Use Internet Elsewhere • The Number of Offline Households Whose Main Reason for Being Offline Is Privacy and Security Reasons • The Number of Offline Households Whose Main Reason for Being Offline Is Not Available Where They Live

continued on next page

Table 16.1 (continued)

Metrics of Interest
Types of Internet Service: Universe: Households With Home Internet Use, Unless Otherwise Stated Below
• The Number of Households That Has At Least One Person to Use Mobile Data for Internet Access • Universe: Households With At Least One Internet User from Any Location • The Number of Households That Use Wired High Speed Technology to Access the Internet • The Number of Households That Use Satellite Technology to Access the Internet • The Number of Households That Use Dialup Technology to Access the Internet • The Number of Households That Buy Home Internet from A Company • The Number of Households That Buy Home Internet from A Public Agency, Non-Profit or Corporation • Number Of Households Where Home Internet Is Provided by Building/Condo, And Included Housing Costs • Number Of Households That Have Home Internet Publicly at No Charge
Online Activities: Universe: Civilians ages 15+ that use the internet
• The Number of Civilians Who Use the Internet for Email • The Number of Civilians Who Use the Internet to Text, or Instant Message • The Number of Civilians Who Use Social Networking • The Number of Civilians Who Publish or Upload Blog Posts, Videos, or Other Content • The Number of Civilians Who Use the Internet for Voice/Video Calls or Conferences • The Number of Civilians That Use the Internet to Watch Videos Online • The Number of Civilians That Use the Internet to Stream/Download Music, Radio, Or Podcasts • The Number of Civilians That Work Remotely • The Number of Civilians That Use the Internet to Look for A Job • The Number of Civilians That Take Online Classes or Participate in Online Job Training • The Number of Civilians That Use Online Financial Services (Banking, Investing, Paying Bills) • The Number of Civilians That Shop, Make Reservations, Or Use Other Consumer Services on The Internet • The Number of Civilians That Sell Goods Over the Internet • The Number of Civilians That Offer Services for Sale Over the Internet • The Number of Civilians That Interact with Household Devices That Use the Internet

16.2.3 Legislation data

The legislation data in this project is from a single source: Pew Charitable Trusts (2021), which is an independent non-governmental organization with a stated mission of "improving public policy, informing the public, and

Table 16.2 Policy Datapoint: An example of what policy data are read in as.

Policy Header	2019 Md. Laws, Chap. 14
Code Title	Local Government Infrastructure Fund (Fund - Broadband)
Year	2019
State	Maryland
State Code Text	Local Government Infrastructure Fund. Provide funds to provide grants and loans to local governments and private providers for improvements to broadband Internet access, provided that the Office of Rural Broadband shall award grants and loans to local governments in a competitively and technologically neutral fashion, provide for fixed and mobile broadband, and target funds to unserved and underserved areas of the State. Further provided that grants and loans may be used for all necessary capital expenses associated with construction or upgrading broadband networks, including but not limited to switches, transmitters, equipment shelters, transport, routers, access points, or network interface devices. Funds shall not be used for operating expenses, including but not limited to leases and customer devices such as handsets, laptops, and tablets . . . $9,680,000

invigorating civic life" (Pew Charitable Trusts, 2021). This source compiles data on public policies that are separable by regulatory discipline; in the case of AIM, policies that impact technology usage were chosen to be the subject. The data downloaded from The Pew Trusts site include the header, the code title, the year the code was passed, the respective state, the state code, and summaries written by Pew researchers. To make the system as reproducible as possible, the summaries were excluded from the data. See Table 16.2.

The data has dimensions including official policy header, the title of the code, the year the policy was enacted, the state the policy was enacted in, and the complete text of the state code. The state code data are the direct wording of each policy. This information is rich with keywords and motivations, which would be a significant challenge for natural language processing (NLP) to make serious inference from, so we chose to use a simple approach to abstracting the text data into a numeric feature that could be used as an input to the model. This is accomplished using term frequency inverse document frequency (TFIDF) and spectral clustering to put

laws into clusters based on similarities in the text of the state code. This approach is described in more detail in Section 16.3.2 of this chapter.

16.2.4 Environmental descriptors

Environmental descriptors are a way to enrich the physical location information encoded in the data. By adding numeric features that describe the environment, hence the name environmental descriptors, the model learns more than just the name of the state that the law is passed in. When dealing with technology such as broadband, for example, an important consideration to lawmakers would be the size of a state. The policies that work for Maryland and Delaware likely wouldn't have the same impact on states such as Texas and Montana, simply because there are very different considerations for that technology in states with lower population densities and more funding per person towards technology initiatives. State funds and federal funds of technology research were one environmental descriptor that we chose to use, because the amount of funding a state has for tech research and development from the government likely differentiates states that have good infrastructure from those that don't. Table 16.3 lists all of the environmental descriptors collected for this project, most of which are used to describe the socioeconomic properties of the state, but some variables have political significance, such as the number of legislators which serve to inform the political side of the problem.

Environmental descriptors provide a means to address the concept of dark data, discussed many times throughout this chapter. By filling in the dataset with the added dimensionality of environmental descriptors, a more complete and detailed dataset provides researchers with more possible trends to correlate with metrics of interest, increases the breadth of data trends for ML algorithms to identify, and makes the predictions of the AIM system less dependent on a single source or type of data, thus increasing the robustness of the algorithm.

16.3 Feature creation

In the case of policies considered in this study, raw data are a string containing a state, year, title, and state code text. In the case of technology

Table 16.3 List of Environmental Descriptors: Tracking physical and socioeconomic descriptors of the states that are used to inform the AIM algorithm on multiple aspects that can be used to compare states.

Environmental Descriptors
• Population
• The Amount of Funding for State Government
• The Amount of Funding State Government Received from Federal Government
• Total Funding (Sum of Federal and State)
• Per Capita Funding (Total Funding Over Population)
• Monthly Unemployment Rate
• Median Household Income
• Gross Domestic Product by State (Millions of Current Dollars)
• Gross Domestic Product Per Capita By State
• Percent Of 4-Year-olds Enrolled in State-Funded Pre-K
• Percent Of People 25 Years and Over Who Have Completed High School (Includes Equivalency)
• Percent Of People 25 Years and Over Who Have Completed a Bachelor's Degree
• Percent Of All Bridges Structurally Deficient
• Total Number of Traffic Fatalities
• Total Renewable Energy Net Generation (Thousand Kilowatt-hours)
• Violent Crime Rate Per 100,000 Population
• Total Expenditures for Public Elementary and Secondary Education (In Thousands of Dollars)
• Degree Granting Higher Education Institutions: All Public Institutions
• Elementary And Secondary Education Expenditures (In Millions of Dollars): Total
• Alternative Fuel Vehicles in Use
• State Intergovernmental Expenditures: Total (In Thousands of Dollars)
• State Intergovernmental Revenue: Total (in Thousands of Dollars)
• Federal Government Expenditures: Total (in Millions of Dollars)
• State Government Tax Revenue: Total (in Thousands of Dollars)
• State Debt Outstanding at End of Fiscal Year (in thousands of dollars): Total
• State Statistics: Number of Representatives in Congress
• Number of Legislators: Total in Senate
• Voting Statistics for Presidential Elections: Number Voting
• Top 1% Income Share

metrics, raw data are web-based tables of survey numbers, counts, dollar amounts, and percentages all separated by state and year. To convert these data into a format that can be mathematically manipulated and made sense of by a machine learning algorithm, several transformations need to first

occur. Data need to be made numerical and standardized, text needs to be processed using natural language processing, and connections need to be made around the central axis of the dataset; in this case, space and time. Whether the variable is state code, population, or the number of households using wired internet, all of the data are bound to a state and year from which it was recorded. This allows us to manipulate the data in such a way that is intuitive from an interpretation point of view.

16.3.1 Policies as data

One manipulation that may not be immediately obvious is "*how do you represent a piece of legislation numerically?*" There are not many examples of this exact problem available in literature. However, there is a very vast knowledge base of NLP tools that have been developed over the past 10 years. Using any single method would be difficult to justify for the purpose of evaluating policies, so using the principle of Occam's razor, a very simple method was chosen. To keep the AIM pipeline modular and adaptable for future developments, the preprocessing of data through NLP and TFIDF can be removed from the pipeline and replaced with other methods of representing the policies in the future.

16.3.2 NLP in AIM

To represent the laws and compare them to one another, numerical representation of text is needed before any further processing can be performed. Natural language processing is a growing and demanding field of AI, so to keep the focus of this project on the AIM system for policymaking, term frequency inverse document frequency (TFIDF) was selected as a straightforward implementation method for transforming the textual data to numeric data. TFIDF is a text vectorization method that is commonly used in machine learning to convert text into a numerical representation. There are two parts to the TFIDF algorithm: the first (Eq. (16.1)) is the term frequency component, where a count of word appearances in the body of text is calculated, and because there are going to be documents of varying length, the count is normalized by dividing all counts by the length of the document. The second (Eq. (16.2)) is the inverse document frequency, or how

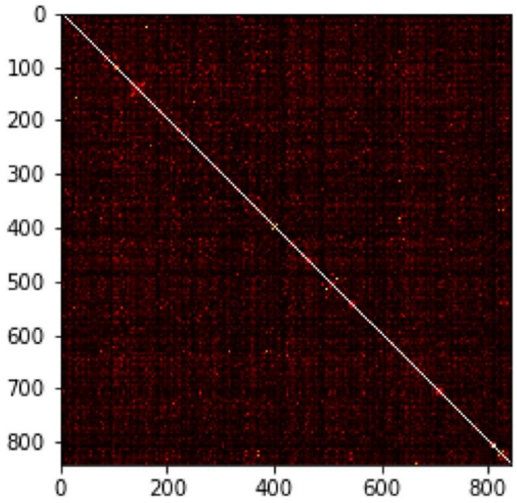


FIGURE 16.4 Similarity Plot: Cosine similarity of TFIDF vectorizations of state codes.

rare a word is in the set of documents. Putting both parts together results in the TFIDF algorithm (Eq. (16.3)), and outputs a vector representation of any body of text.

$$tf\,(t,d) = \log(1 + freq\,(t,d)) \tag{16.1}$$

$$idf\,(t,D) = \log(\frac{N}{count\,(d \in D : t \in d)}) \tag{16.2}$$

$$tfidf\,(t,d,D) = tf\,(t,d)\,.idf(t,D) \tag{16.3}$$

The TFIDF vectorization for each policy was calculated using the equations above, and then a similarity matrix was calculated based on the vector representations of each policy using cosine similarity. Similarity values range from 0 to 1, with 1 meaning identical and 0 meaning no similarity. By visualizing the matrix as a heat map in Fig. 16.4, the entire matrix with all 693,889 connections is presented with a line of 1's for self-similarity and other highs and lows in the plot. There is no discernable pattern, because the policies are sorted alphabetically by state, which is identifiably not associated with the content of the laws.

The similarity matrix is the next step in converting the laws into a meaningful clustered representation that can be used as an input to the neural network.

Table 16.4 Clustering Sizes: Laws grouped by TFIDF vectorization similarities.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Size (n)	93	74	145	59	44	63	40	83	77	55

16.3.3 Spectral clustering of laws

Using the similarity matrix visualized in Fig. 16.4, similarity values can be transformed into clusters using a variety of techniques, one of which is called spectral clustering. Spectral clustering in general encompasses a class of clustering algorithms that produce high-quality clusters on small datasets. With a computational complexity of $O(n^3)$, the dataset that this project is concerned with is small enough to be practical, with less than 1000 data points. However, expanding by orders of magnitude in the future may require a rethinking of this procedure (Yan et al., 2009). With a selected number of clusters equal to 10, very even groups were able to be generated as seen in Table 16.4. This number of groups was also ideal, because converting into a one-hot encoded format and using that as the input to the neural network was not a cumbersome process.

16.3.4 Technology usage as data

To represent changes in the technology metrics that would be useful to predict, a first-order derivative labeling schema was employed to transform the data from static points to velocity indicators. For a given year i, the metric value x_i is used as a baseline, and the next valid year is used to find the difference. If the value in the next valid year sees an increase in the metric, then the function returns the positive difference divided by the number of years, indicating an upward trend and a positive slope. If the value in the next valid year sees a decrease in the metric, then the function returns the negative difference divided by the number of years, indicating a downward trend and a negative slope.

$$f\,(\mathrm{X}, i) = \frac{x_{i+\Delta i} - x_i}{\Delta i} \tag{16.4}$$

The resulting data are used as labels for the neural network to train on and predict for new data points. By predicting changes in the metrics of interest, AIM is able to predict the associated impact on each of the 47 metrics. The word associated is chosen, because there is no definite correlation or causation between the change in the metric and the passing of the law, and the word impact is chosen, because the change in the metric is a representation of the impact on the lives of constituents who were surveyed or studied as sources of the data.

16.4 Learning the trends

As an AI-enabled system, the learning method is very important and was carefully chosen for a project involving a topic as sensitive and susceptible to criticism as political advocacy. A few of the criteria for the algorithm that would learn the trends were the following: it must be explainable, it must be transparent (as opposed to black box), and it must be accurate in making predictions.

16.4.1 Neural network predicting AIMs

Neural networks are biologically inspired mathematical functions that have individual layers of nodes, each node in a layer being connected to every node in the subsequent layer with learned weighted connections and used to generate output values (of any dimension) based on input values (of any dimension). This last property was particularly important to this project, having 47 output values and 40 input values, because not all algorithms or structures of neural networks are as flexible.

The network used for AIM has 3 hidden layers, all of which are linear, with varying numbers of nodes in each layer. The objective of predicting the associated impact measures is fulfilled by training and deploying the model shown in Fig. 16.5. Referring to the pipeline in Fig. 16.2, the model is the endpoint of the AIM system backend and the starting point for the AIM system frontend, where new laws can be predicted upon to assist policy researchers and data scientists.

Fig. 16.5 is a visual representation of the ANN used in the AIM pipeline. The color and opacity of connections between nodes are reflective of the

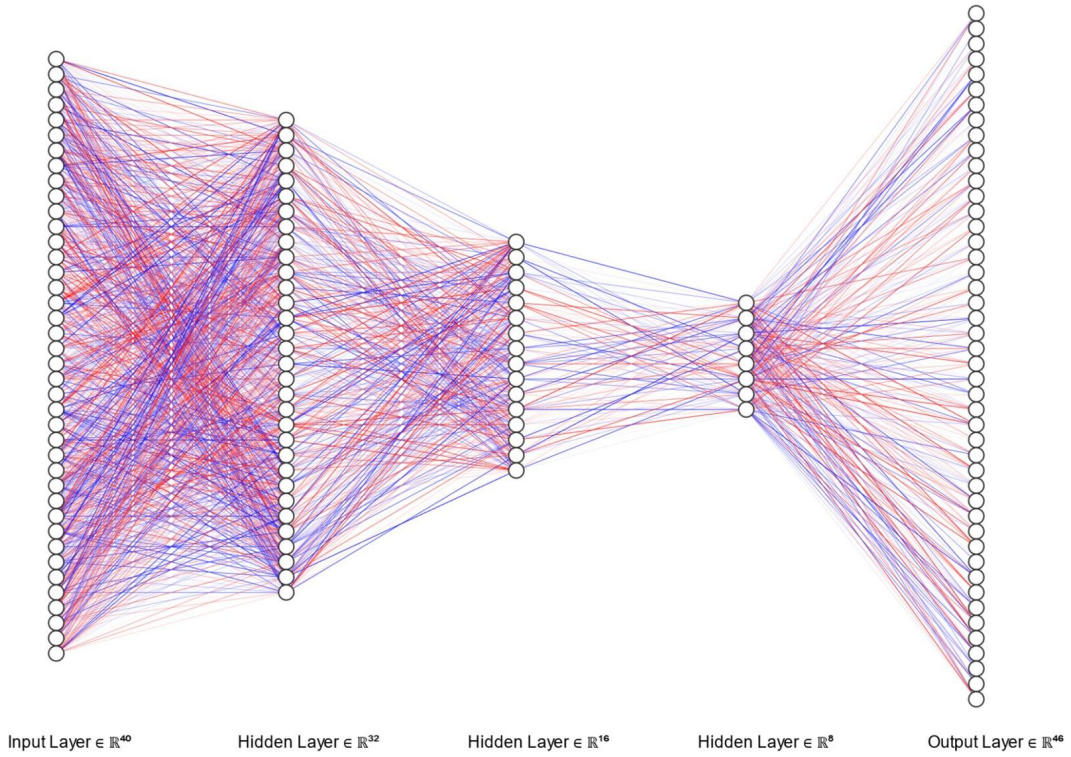


FIGURE 16.5 Artificial Neural Network: Visualization of weights and layers in the neural network.

weight generated using an open-source online tool. Creating visual representations of neural networks (LeNail, 2019) is a step towards greater explainability in AI, one of the core pillars of AI assurance.

16.4.2 Training metrics

Training a neural network requires a number of decisions to be made of the programmer building the algorithm. The first is the loss function; this value is what the algorithm is seeking to minimize and will train accordingly to produce the lowest allowable loss value. In AIM we implement the mean squared error loss function, which adheres to the format of Eqs. (16.5) and (16.6).

$$\ell(x, y) = mean(L) \tag{16.5}$$

$$L = \{l_1, \ldots, l_N\}^\top, \quad l_n = (x_n - y_n)^2 \tag{16.6}$$

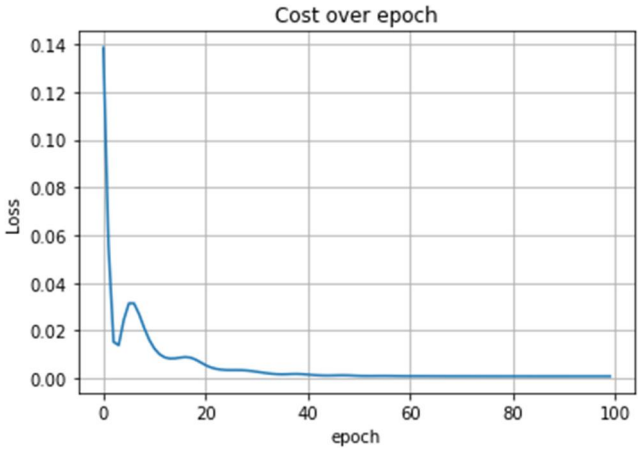


FIGURE 16.6 Model Evaluation Plot: High cost initially, with training the model becomes more accurate.

Observing the loss as training proceeds in Fig. 16.6, the algorithm performs well, beginning relatively high for normalized data at 0.14 (which is expected) and decreasing drastically and consistently after training multiple epochs.

With a train-test split of 80% train and 20% test, the data are segregated and ready to evaluate. The final loss for the training dataset was 0.0009, and the loss for the test set was 0.0036. As seen in Fig. 16.6, the cost over epoch plot of training the neural network shows a major improvement (decrease) in loss over the first 50 epochs before leveling out near 0. These results provide high confidence in the ability for the model to accurately predict the associated impact on metrics of interest, the primary objective of AIM.

16.4.3 Prediction results

For an upcoming piece of legislation, the input to AIM consists of a state, a year, and a state code. Through the processing established in the system, environmental descriptors are automatically appended, a cluster is assigned based on the text of the policy, and the trained neural network predicts the change in the technology metrics. Below is an example of what the system outputs given a piece of legislation; importantly, the state code itself is used to assign a data point to a similarity cluster and not any third-

Table 16.5 Colorado Legislation Input: Datapoint from the test set of the data, "Colo. Rev. Stat. 29-27.401" shows the typical input to the system.

Policy Header	Colo. Rev. Stat. 29-27-401
Code Title	Legislative declaration
State	Colorado
Year	2017
State Code Text	(1) The general assembly finds and declares that: (a) The permitting, construction, modification, maintenance, and operation of broadband facilities are critical to ensuring that all citizens in the state have true access to advanced technology and information; (b) These facilities are critical to ensuring that businesses and schools throughout the state remain competitive in the global economy; and (c) The permitting, construction, modification, maintenance, and operation of these facilities, to the extent specifically addressed in this part 4, are declared to be matters of statewide concern and interest. (2) The general assembly further finds and declares that: (a) Small cell facilities often may be deployed most effectively in the public rights-of-way; and (b) Access to local government structures is essential to the construction and maintenance of wireless service facilities or broadband facilities.

party generated summaries, making the system more robust and generally applicable.

With the prediction results displayed as a table, interpretation of the predictions can be made by an expert as to whether the results reflect positively or negatively on the proposed law. What Table 16.6 shows is annualized changes in the individual metrics. So, for instance, the law in Table 16.5 specifically states that building internet infrastructure is critical to the mission of the state and making Colorado competitive in a global economy. It would benefit researchers to know what the expected impact would be associated with a law like this regarding broadband access and internet accessibility. By looking at the variables that correspond to trends related to broadband access and internet accessibility, we can see what AIM delivers and how it can be used to inform policy researchers.

The results from Table 16.6 describe that Colorado citizens with internet at their home is predicted to increase by 196,542, and the number of citizens who don't have internet at home because it is too expensive decreases by 45,220. This example uses a law that is from the test set. However, the sys-

Table 16.6 Predicted Metrics: AIM produces a prediction for every metric by using the input in Table 16.5 and the trained neural network from Fig. 16.5.

Metrics of Interest	Predictions for "Colo. Rev. Stat. 29-27-401"	Metrics of Interest (2)	Predictions for "Colo. Rev. Stat. 29-27-401" (2)
desktop_use	+13589.0	useIntElsewhere	+3069.0
laptop_use	+275224.0	unavailableInt	+1115.0
tablet_use	+42265.5	mobileDataUsers	+168515.0
mobile_use	+374395.5	wiredHigh-SpeedUsers	+212822.0
smartTV_use	+292803.0	satelliteUsers	-25186.5
wearable_use	+225967.0	dialUpUsers	-13540.0
intUsers_above3	+440629.5	intPrivateISP	+190382.5
intUsers_above15	+358238.0	intPublicISP	+2152.5
homeIntUsers	+467141.5	intIncluded	+15166.0
workIntUsers	+150907.0	intPublicFree	+2028.0
schoolIntUsers	+72093.0	emailUsers	+276797.0
cafeIntUsers	+106189.5	textIMUsers	+293107.5
altHomeIntUsers	+181263.0	socialNetUsers	+147679.0
travelIntUsers	+371278.5	publishUsers	+111882.0
publicIntUsers	+51240.0	onlineConfUsers	+152372.0
anyHomeIntUsers	+16158.5	videoUsers	+199706.5
intAtHome	+196542.0	teleworkUsers	+118745.5
noIntAtHome	-155053.5	jobSearchUsers	+27160.5
homeEverOnline	-67682.5	onlineClassUsers	+18923.0
noNeedInt	-74163.5	financeUsers	+217197.5
noExpensiveInt	-45220.0	eCommerceUsers	+257499.5
noComputerInt	-10388.5	sellingGoodsUsers	-4515.0
noPrivSecInt	+1680.5	iotUsers	+191678.5

tem could perform this prediction for laws that are currently in the writing process to evaluate potential effectiveness. This is a perfect example of how policies can be associated with impacts on metrics of interest, and though the relationship is not causal or even measurably correlated, the trends can still be used to inform policymakers in directing their investigations, while performing policy monitoring and policy advocacy activities.

16.5 Discussions and future directions

The AIM system is a proof of concept that AI-enabled tools, in conjunction with big data and cloud computing services, are ready to enter the legislative arena. Methods presented in this chapter demonstrate the ability to collect and utilize different types of data, unify them into a single format, and make predictions with a neural network for the purpose of making the legislative process better informed. Addressing dark data is a primary goal of this project, and in doing so we found that though no artificial system will be as naturally intuitive as a human policy researcher, it can be extremely useful at delivering accurate measurements very quickly, and in political situations that the United States finds itself in today, accuracy, transparency, and accountability, are very valuable characteristics.

16.5.1 Feasible applications

As mentioned, many times in this chapter, AIM is a tool that is meant to assist lawmakers and policy researchers in doing their due diligence when it comes to evaluating technology policies. Should it be deemed necessary, changing the datasets to be more inclusive of domains other than technology, for example, economics or transportation or foreign policy, this application architecture could be applied to much more complicated and critical areas of politics. One issue with raising the stakes of the AIM system is that the algorithm needs to be more accurate, more robust, and much less susceptible to making mistakes. There are many methods described in this book that can be used to achieve all of these properties. AI assurance is a critical domain for computer science, mathematics, cybersecurity, and with this application, politics. There is no doubt that as the computational domain exits the control of its human operators in the future; artificially intelligent systems will be what dictates many areas of decision-making. Therefore it is of the utmost importance that AI now is developed in a safe, explainable, predictable, and human-centric fashion.

16.5.2 Future directions

Many aspects of the AIM architecture deserve to be scrutinized and improved upon. For example, the environmental descriptors used in the data

could be expanded upon to make the feature set of the ML algorithm more detailed and less ignorant of dark data. It will never be possible to be fully inclusive of all information that exists, but the pursuit of that inclusivity will drive the improvement of accuracy and efficiency of AI systems for years to come. With constant advancements in the development of AI algorithms specifically; it is also undoubtedly true that newer and more computationally extensive methods of learning the trends in the data could in the future replace the neural network used for this system.

16.6 Ethics of AI in public policy

Ethics is a monumental consideration when incorporating any technology into a human-centric process. This is no different (from an AI assurance perspective) when making robots in a factory than it is when making inferences on big datasets for policymaking. This section discusses some of the ethical considerations surrounding the implementation of AI in policy research and what can be done to ensure safe, predictable, and useful results.

16.6.1 Data in the legislative process

With the combination of big data being collected and AI used to interpret it, the influence of intelligent algorithms will only grow (McNeely and Hahm, 2014). The portion of the legislative process that AI seems best fit to improve is the research and evaluation processes. These make up a significant amount of the research that is dedicated to policy advocacy—an ongoing process for any piece of legislation—and therefore any benefits from AI-enabled systems, such as AIM would be quickly identifiable and measurable. This kind of implementation, acting more as an AI-assist, offers much less risk than fully automated AI applications with the same fundamental operations underneath (Pencheva et al., 2020). With AI-assisted processes, the benefits of AI integration can be observed without compromising the integrity or interpretability of the outcomes.

16.6.2 AI and bias

Bias in the legislative process is an extremely serious concern that transcends mathematics and algorithms and enters the domain of sociology

and history. There are examples of geographic, racial, age, and gender-based bias across all subjects of law (Levinson and Smith, 2012; Peller, 1992) and to protect the progress that has been made to creating inclusive and constitutionally sound laws in the US, the introduction of AI algorithms must be ensured not to return to a legal system riddled with discrimination and prejudice. Bias in AI is defined mathematically in the context of different algorithms or processes; however, we also consider bias in terms of results in predictions and inference on different groups of people. One way that analytical tools in general have become a source for bias in legal systems is through risk assessment. There is a tendency for actuarial risk assessment to produce a "ratchet effect" on members of high-risk categories, with detrimental effects on employment, educational, familial, and social outcomes; also see Harcourt (2015). The use of mathematics and technology for evaluating people in the past has failed to overcome the lingering grasp of racism for instance, and it would be shortsighted to think that AI could inherently be immune to that same fate. Bias of all kinds is one of the greatest concerns when using AI-enabled technologies for public policy advocacy, because the goal is to produce legislation that better governs the people, and governments that better represent the constituents.

16.6.3 AI assurance and the law

This chapter discusses the use of AI to better inform the process of creating new policies and laws. Though the topic of AI for policy is not a rare one, it typically is in the manner of creating regulations for AI systems themselves, and not the other way around. It is important to develop, implement, and introduce AI in ways that are safe and responsible, and with this in mind, governments have recently started to define AI legally and begun to lay the foundations for future regulations on the technology.

The European Union has become one of the first government bodies to propose a regulatory framework on the use and deployment of artificially intelligent systems (Artificial Intelligence Act, 2021). The proposal is fundamentally broad as any foundational policy must be to allow for development and innovation to continue. To establish the beginning of the framework, the proposal distinguishes between low-risk, high-risk, and un-

acceptable risks to the safety of AI users. Examples of unacceptable risk to AI users are systems that could "manipulate vulnerabilities of specific groups of people" and give them a "social score." There are already systems that are in place to measure, track, and evaluate people in a manner as described in this risk category so it is a very important topic to address concerning AI (Wong and Dobson, 2019). AI applications must be distinguished from high-risk to unacceptable, because some of the most powerful uses for AI could be very risky and require strict oversight, however, should not be unacceptable. Not all AI applications are as dystopian, though. The proposal expands to infrastructure implementations and critical services defining them as high-risk and requires rigorous documentation and risk management for all companies that use AI in these areas. Then low-risk applications, such as photo-filters, emotion detection, all must comply with transparency obligations so that users know when they are interacting with AI-based technology (Yaros et al., 2021).

A system of legal advocacy based on AI interpretations should be classified as low-risk when there is still a human filter between the results and the legislative decision. If that filter is to be removed, it would be certainly a high-risk application that requires trustworthiness, security, explainability, and above all, assurance.

Through the use of methods presented in this book, AI algorithms can be made more fair, trustworthy, explainable, ethical, safe, and secure while maintaining their ability to inform and improve processes and systems in all domains.

References

Artificial Intelligence Act, 2021. (2021) 206: Proposal for a Regulation of the European Parliament on AI. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

Batarseh, F.A., Freeman, L., Huang, C.H., 2021. A survey on artificial intelligence assurance. Journal of Big Data 8, 60. https://doi.org/10.1186/s40537-021-00445-7.

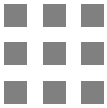
Gordon, I., Lewis, J., Young, K., 1977. Perspectives on policy analysis. Public Administration Bulletin 25, 26–35.

Grimm, D.J., 2018. The dark data quandary. American University Law Review 68, 761.

Harcourt, B., 2015. Risk as a Proxy for Race: The Dangers of Risk Assessment. 27 FED. SENT'G REP. 237. https://scholarship.law.columbia.edu/faculty_scholarship/2564.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Library Trends 57, 280–299. https://doi.org/10.1353/lib.0.0036.
Pew Charitable Trusts, 2021. How we work. https://pew.org/2jI2cMJ.
Keating, Bill, 2016. The legislative process. https://keating.house.gov/policy-work/legislative-process.
Klein, J.T., 1990. Applying interdisciplinary models to design, planning, and policy-making. Knowledge, Technology & Policy 3, 29–55. https://doi.org/10.1007/BF02736654.
LeNail, 2019. NN-SVG: publication-ready neural network architecture schematics. Journal of Open Source Software 4 (33), 747. https://doi.org/10.21105/joss.00747.
Leslie, D., Burr, C., Aitken, M., Cowls, J., Briggs, M., Jones, L.T.C., 2021. Artificial intelligence, human rights, democracy, and the rule of law: a primer. Zenodo. https://doi.org/10.5281/zenodo.4639743.
Levinson, J.D., Smith, R.J., 2012. Implicit Racial Bias Across the Law. Cambridge University Press.
McNeely, C.L., Hahm, J., 2014. The big (data) bang: policy, prospects, and challenges. Review of Policy Research 31, 304–310. https://doi.org/10.1111/ropr.12082.
Nutley, S., Davies, H., Walter, I., 2002. Evidence Based Policy and Practice: Cross Sector Lessons from the UK.
O'Mathúna, D., Næsager, Lene., Greubel, J., 2021. Public trust and evidence-based policymaking: Lessons from the COVID-19 response. European Policy Center.
Peller, G., 1992. Criminal law, race, and the ideology of bias: transcending the critical tools of the sixties. Tulane Law Review 67, 2231.
Pencheva, I., Esteve, M., Mikhaylov, S.J., 2020. Big data and AI – a transformational shift for government: so, what next for research? Public Policy and Administration 35, 24–44. https://doi.org/10.1177/0952076718780537.
Segone, M., 2008. Evidence-based policy making and the role of monitoring and evaluation within the new aid environment. In: The Role of Monitoring and Evaluation in Evidence-Based Policy Making, p. 16.
Thierer, A., O'Sullivan, A.C., Russell, R., 2019. Artificial Intelligence and Public Policy 56.
Wong, K.L.X., Dobson, A.S., 2019. We're just data: exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies. Global Media and China 4, 220–232. https://doi.org/10.1177/2059436419856090.
Yan, D., Huang, L., Jordan, M.I., 2009. Fast approximate spectral clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09. Association for Computing Machinery, New York, NY, USA, pp. 907–916.
Yaros, O., Bruder, A., Hajda, O., Graham, E., 2021. The European Union Proposes New Legal Framework for Artificial Intelligence | Perspectives & Events | Mayer Brown. https://www.mayerbrown.com/en/perspectives-events/publications/2021/05/the-european-union-proposes-new-legal-framework-for-artificial-intelligence.

This page intentionally left blank



Index

B

C

D

E

F

G

N

O

P

Q

R

S

T

AI Assurance

Towards Trustworthy, Explainable, Safe, and Ethical AI

By Feras A. Batarseh and Laura J. Freeman

Provides a leading-edge guide to AI assurance to facilitate developing and applying AI in a trustworthy, explainable, fair, safe, secure, and ethical manner

Key Features

- Provides readers with in-depth understanding of how to develop and apply Artificial Intelligence in a valid, explainable, fair, and ethical manner.
- Includes description of providing assurance to various AI methods, including deep learning, machine learning, reinforcement learning, computer vision, agent-based systems, natural language processing, text mining, predictive analytics, prescriptive analytics, knowledge-based systems, and evolutionary algorithms.
- Presents techniques for efficient and secure development of intelligent systems in a variety of domains, such as agriculture, healthcare, cyber security, government, and education.
- Provides readers with complete example datasets and code associated with the methods and algorithms developed in the book.

AI Assurance provides readers with solutions and foundational understanding of the methods that can be applied to test AI systems and provide assurance. Anyone developing software systems with intelligence, building learning algorithms, or deploying AI to a domain-specific problem (such as allocating cyber breaches, analyzing causation at a smart farm, reducing readmissions at a hospital, ensuring soldiers' safety in the battlefield, or predicting exports of one country to another) will benefit from the methods presented in this book.

AI assurance is now a major piece in AI and engineering research, and this book serves as a guide to researchers, scientists, and students in their studies and experimentation with AI. Moreover, as AI is being increasingly discussed and utilized at government and policymaking venues, the assurance of AI systems—as presented in this book—is at the nexus of such debates.

ACADEMIC PRESS
An imprint of Elsevier
elsevier.com/books-and-journals