Centre for International
Governance Innovation

CIGI Papers No. 296 — July 2024

Explainable AI Policy
It Is Time to Challenge Post Hoc Explanations

Mardi Witzel, Gaston H. Gonnet and Tim Snider

Centre for International
Governance Innovation

CIGI Papers No. 296 — July 2024

Explainable AI Policy

It Is Time to Challenge Post Hoc Explanations

Mardi Witzel, Gaston H. Gonnet and Tim Snider

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Centre for International Governance Innovation or its Board of Directors.

For publications enquiries, please contact publications@cigionline.org.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Credits

Research Director, Transformative Technologies Tracey Forrest
Director, Program Management Dianna English
Program Manager Erin Chreptyk
Senior Publications Editor Jennifer Goyder
Publications Editor Christine Robertson
Graphic Designer Abhilasha Dewan

Table of Contents

About the Authors

Mardi Witzel is the CEO and co-founder of PolyML, an early-stage Canadian artificial intelligence (AI) and machine learning company with novel technology. In her role, Mardi is responsible for setting the organization's strategy and executing the growth plan. Mardi is recognized as an expert in the field of AI governance, serving as a member of the province of Ontario's Expert Working Group for a Trustworthy AI Framework and a member of the AI Ethics Advisory Panel for the Digital Governance Council of Canada.

A board director with 20 years' experience in board governance, stakeholder engagement and strategic planning, Mardi also writes on policy issues relating to AI, ESG (environmental, social, governance) and the challenges of innovation and sustainability in the digital economy. She is currently serving a second term as a public representative on the Chartered Professional Accountants Ontario (CPA Ontario) Council and previously served as chair of the Muskoka Watershed Advisory Group and chair, KidsAbility Foundation Board.

Gaston H. Gonnet is PolyML's chief data scientist. He received his doctorate in computer science from the University of Waterloo. He is skilled in symbolic and algebraic computation, in particular, solving equations (symbolically and numerically), system development, limit and series computation, heuristic algorithms, text searching and sorting algorithms.

Gaston also developed Darwin Programming Language for biosciences, which would become the basis for OMA, a package for gene orthology prediction.

Tim Snider is PolyML's chief technology officer. He holds a master of computer science from McMaster University and has a Bouvier-poodle named Patsy.

Tim has worked with big data since helping to move the Oxford English Dictionary from paper to disk and creating one of the first (pre-Google) full-text indexes of the internet. Tim has also worked on a variety of commercial and research projects, including a column-oriented analytical SQL engine, a SQL to C translator and a text analysis tool based on probabilistic finite automata.

Acronyms and Abbreviations

AI	artificial intelligence
AIA	Algorithmic Impact Assessment
AIDA	Artificial Intelligence and Data Act
Fiins	feature importance insights
HLEG	High-Level Expert Group
ICO	Information Commissioner's Office
LLMs	large language models
NIST	National Institute of Standards and Technology
LIME	local interpretable model-agnostic explanations
SHAP	SHapley Additive exPlanations

Executive Summary

The focus of this paper is on policy guidance around explainable artificial intelligence (AI) or the ability to understand how AI models arrive at their outcomes. Explainability matters in human terms because it facilitates including an individual's "right to explanation" and it also plays a role in enabling technical evaluation of AI systems. In respect to both, explainability may be viewed as a sort of "first among equals" — functioning as a principle for trustworthy AI that affords visibility into performance on measures associated with a wide range of other principles relating to the behaviour of AI systems.

The paper begins with an examination of the meaning of *explainability*, concluding that the constellation of related terms serves to frustrate and confuse policy initiatives. Following a brief review of contemporary policy guidance, it argues that there is a need for greater clarity and context-specific guidance, highlighting the need to distinguish between ante hoc and post hoc explainability, especially in high-risk, high-impact contexts.

The question of whether post hoc or ante hoc methods have been employed is a fundamental and often-overlooked question in policy. The paper argues that the question of which method should be employed in a given context, along with the requirement for human-level understanding, is a key challenge that policy makers need to address. A taxonomy for how explainability can be operationalized in AI policy is proposed and a series of recommendations is set forth.

The paper also includes two appendices for readers with more technical backgrounds: one containing a review of commonly used explainable AI methods and the other showcasing the example of a novel technique for the development of explainable AI models.

Introduction

If 2022 was the year AI broke through to the masses, then 2023 was the year it made its way into the corridors of power, with governments all over the world waking up to the urgency of regulating it. While AI in the form of machine learning has been around for decades, the landscape changed in November 2022 with the release of ChatGPT. AI became one of the most discussed topics in the world during 2023.

The Organisation for Economic Co-operation and Development logs a live repository[1] of more than 1,000 AI policy initiatives from more than 69 countries, territories and the European Union. The European Union, Canada, the United States and the United Kingdom all have AI-related policy instruments that are either enacted, under review or in development, and many of these embrace a common, or at least overlapping, set of principles, including transparency and explainability, accountability, fairness and bias, privacy, safety, security, technical robustness and human-centredness.

The focus of this paper is on explainable AI (sometimes referred to as XAI), and more specifically the consideration of explainability from a policy perspective. Explainable AI refers to the ability of humans to interpret and understand how AI models arrive at their decisions or outputs. Explainable AI stands in contrast to AI black boxes, or models where it is not possible to gain insight into the mechanisms by which they translate inputs into outputs. While numerous academic and technical papers have been written on the topic of explainable AI, much less has been written about it from a policy perspective. Luca Nannini, Agathe Balayn and Adam L. Smith (2023) provide an excellent survey on the subject, highlighting the dearth of actionable regulatory standards, the coarseness of technical specifications in proposed legislation and nascence of the concept of explainability in research, essentially the oversimplification of a novel and complex problem.

The question of why policy makers, including civil servants who inform the legislative process and regulators who implement the rules, have not been more demanding of genuine explainability

[1] See https://oecd.ai/en/dashboards/overview.

in AI systems, is curious. Consider the large language models (LLMs) that are emerging like rapid fire today, and the lack of insight into anything much about them, including what data they were trained on and how they work. Why is nobody challenging this?

There are numerous reasons why people in general should care about explainability in AI, but ultimately it is about trustworthiness and the role that explainability plays in that assessment. The EU High-Level Expert Group (HLEG)[2] stipulated that trustworthy AI must satisfy three necessary conditions: compliance with applicable laws and regulations (lawfulness), adherence to ethical principles and values (ethics) and safety, security and reliability (robustness). There is acknowledgement of the difficulty in ensuring that these conditions for trustworthy AI hold, given the lack of proven methods to translate them into practice (Markus, Kors and Rijnbeek 2021). There is, however, broad recognition of the potential for explainability to contribute to the assurance of trustworthy AI, and it is referenced in multiple frameworks and policy initiatives.

There are several reasons why policy makers should care about explainability, including a human's "right to explanation" and the role that explainability plays in facilitating the evaluation of other model attributes and performance. In respect to both, explainability may be viewed as a sort of "first among equals" — essentially functioning as a principle for trustworthy AI that affords visibility into performance on measures associated with all the other principles.

The focus in this paper is on policy guidance pertaining to the explainability of traditional forms of AI, such as machine learning.[3] The authors recognize the central role that generative AI will play across all economic sectors going forward, and the deep-rooted concern over the black box nature of the underlying technology. However, they have elected not to address explainability in generative AI here, owing to the vastly different approach to the collection of data, and the training, deploying and monitoring of foundational models as compared to traditional machine learning.

The paper is divided into three main sections:

- how explainability relating to AI is defined and what it means today;
- how the explainability of AI is currently approached in governance and policy; and
- how the explainability of AI can best be operationalized in policy going forward.

Following the examination and analysis in these sections, the paper offers a set of recommendations. Additionally, there are two appendices:

- a survey and review of methods for achieving explainable AI that should be considered by policy makers in the development of effective policy; and
- an example of best practice for achieving explainable AI with a proposed new category of explainable AI.

How Explainability Is Defined in a Policy Context Today

While there is evidence of an upward trend in the interest in explainable AI (Islam et al. 2022), there has not been a clear and consistent understanding of what is meant by the term. This absence of clarity around definition is not trivial, and policy around explainability in AI can only be effective with it being addressed. It is often said that "you cannot manage what you do not measure." We also need to recognize that you cannot manage (or measure) what you do not define, and in the case of AI explainability, this has implications for a range of stakeholders including developers, deployers, users, subject matter experts, impacted parties and regulators.

[2] The HLEG is an independent group of experts set up by the European Commission. It represents one of the earlier examples of a group tasked with the development of a set of principles for ethical and trustworthy AI, and its work remains influential today because the principles it established are woven into many of the frameworks that have followed.

[3] The terms machine learning and traditional AI are used interchangeably in this paper, each noted to be forms of the subset of AI known as artificial narrow intelligence. Machine learning is a branch of AI, broadly defined as the capability of a computer to imitate intelligent human behaviour using data and algorithms. Machine learning is distinguished from generative AI, which while based on machine-learning techniques, and itself a form of artificial narrow intelligence (at present), carries with it a different set of considerations from technology, general policy and explainability perspectives.

Despite the lack of a universally agreed-to definition, literature from both the academic and policy worlds offers useful guidance. On balance, the consideration in academic papers is deeper and more nuanced than that found in regulatory and policy instruments. This is probably due to the technical nature of the topic, and the practical reality that doing the subject justice requires an appreciation of the data science aspect of both AI and its explainability. This is not to say that policy makers and regulators need to be technical experts, but rather they need to be sufficiently informed about the technology and the methodologies underlying it to provide appropriate guidance.

While there is a loud call for explainability from policy makers that is picked up in mainstream media and consumer circles, this discussion often stays at a high level and neglects to dissect some of the more critical distinctions between different types and levels of explainability or the role of context. For this reason, the paper starts the review of what is meant by explainability in AI by first turning to academic sources.

The Constellation of Terms around Explainability

The academic literature acknowledges an array of explainable AI terms that are often used interchangeably, including transparency, explainability and interpretability, and seeks to distinguish these in a meaningful way. Transparency is more straightforward to tease out as a separate concept, whereas there is a lot of overlap and inconsistency of use between explainability and interpretability. This quibble over semantics should not be perceived to be ivory tower or excessive, but rather this is a case of academic discourse leading the field.

Transparency relates to how AI is used in organizations and the openness of communications around how and where it is being applied (Balasubramaniam et al. 2023). It is more of an organizational than a technical phenomenon, addressing practices around reporting and disclosure. Transparency relates directly to trust and an assurance that people understand the methods of AI in use, the behaviour of AI systems, the where, why and how AI has been used and the difference between AI for decision making and AI as an assistive tool.

Best governance practices around AI transparency include an organization's policies for things such as organizational and employee AI competency; the rationale for AI use; the identification of roles, responsibilities and accountability for AI and its outcomes; the approach to procurement, development and deployment of AI; the practice of inventorying AI by technology type and/or use case; and the organization's approach to measuring its performance over time.

In contrast, explainability is more of an AI system-level concept. At the heart of the contemporary academic discourse is agreement that explainability of AI needs to be both meaningful to a human and accurately reflective of what a model is doing to process inputs and generate outputs (Sokol and Vogt 2023). Beyond the broad acceptance of the human and technical elements underpinning explainability, there is confusion owing to the different terms that are used in relation to these elements.

Aniek F. Markus, Jan A. Kors and Peter R. Rijnbeek (2021) suggest that "interpretability" and "fidelity" are two sub-components of explainability, where interpretability refers to the presence of a single, unambiguous rationale and fidelity describes the entire dynamic of the task model, including the degree to which an explanation is complete and sound. In contrast, Linardatos Pantelis, Vasilis Papastefanopoulos and Sotiris Kotsiantis (2021) consider interpretability to be a broader term than explainability.

In one of the most widely cited papers on the subject, interpretability is conceptualized as a distinct or extended form of explainability (Rudin 2019), where models are "inherently interpretable" based on the transparency attributes of the algorithms used to build them. In contrast, "explainable AI" refers to explanations that are produced for black box models, through the construction of surrogate models that are built to estimate what the task model is doing (Moradi and Samwald 2021; Ali et al. 2023). These two approaches are called ante hoc (that is, inherent interpretability) and post hoc (that is, after-the-fact) explainability, respectively, and the distinction between them, along with the governance implications, is widely discussed in academic literature.

Boiling all this down, the authors conclude that for AI to be explainable, the explanation must address the model process and outcome (that is, the technical dimension of explainability), as well as be understandable to a human (that is, the human dimension of explainability).

Building on the conceptualization of explainability as a set of underlying factors (Markus, Kors and Rijnbeek 2021), the distinction between ante hoc and post hoc explainability can be incorporated and the following simple framework for thinking about and assessing explainability of AI models can be offered.

Figure 1: Framework for Explainability of AI Models

Explainability: basis for assessment	Ante Hoc or Post Hoc	Level of Human Understanding		Level of Technical Accuracy	
		Clarity (Unambiguous)	Parsimony (Simple, concise)	Completeness (Entire process explained)	Soundness (Correct, truthful)

Source: Authors.

The Appearance of Explainability in Policy Today

This section moves beyond definition and covers what the preponderance of existing policy on explainable AI means when referencing "explainability" today, highlighting how confusing it is and the need for more clarity and context-dependent guidance. Globally, policy differs in addressing the technical and human-level dimensions of explainability, as frameworks vary in both the use of terms and the level of detail ascribed to concepts.

In the United States, explainability requirements appear in non-binding guidance from the White House and the National Institute of Standards and Technology (NIST). NIST (2001) has published a set of four principles of explainable AI: explanation, meaningful explanation, explanation accuracy and knowledge limits. The US Blueprint for an AI Bill of Rights[4] specifies a requirement for "notice and explanation" calibrated to the level of risk but with no specific guidance or mechanism for enforcement (Office of Science and Technology Policy 2022). Canada represents more of a mixed bag, where explainability does not appear in Canada's proposed Artificial Intelligence and Data Act (AIDA)[5] but does show up in the AIDA Companion Document[6] with human oversight and monitoring as a regulated activity, and a future requirement for assessing the level of interpretability needed and making the design decisions accordingly.

On balance, there appears to be growing recognition of the need for specificity in both technical and human-focused guidance, as evidenced by the differences in the European Union's handling of the topic, from the 2021 version of the proposed Artificial Intelligence Act[7] to the 2023 publication of the proposed amendments.[8] The topic of explainability makes an appearance in the proposed EU Artificial Intelligence Act in title III, chapter 2, article 13 (Transparency and Provision of Information to Users). In the 2021 version, the terms interpretability and transparency

[4] See www.whitehouse.gov/ostp/ai-bill-of-rights/.

[5] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Sess, 44th Parl, 2022 (first reading 16 June 2022), online: <www.parl.ca/legisinfo/en/bill/44-1/c-27>.

[6] See https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

[7] EC, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* COM(2021) 206 final 2021/0106 (COD), online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.

[8] See www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

are used interchangeably, and no guidance is given with regard to level or type of explainability:

> High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.[9]

In the June 2023 amendments to the act, proposed changes reflect recognition of both the role of audience and system understanding:

> 1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable *providers and* users to *reasonably understand* the system's *functioning*. Appropriate transparency shall be ensured *in accordance with the intended purpose of the AI system*, with a view to achieving compliance with the relevant obligations of the provider *and user* set out in Chapter 3 of this Title.
>
> *Transparency shall thereby mean that, at the time the high-risk AI system is placed on the market, all technical means available in accordance with the generally acknowledged state of art are used to ensure that the AI system's output is interpretable by the provider and the user. The user shall be enabled to understand and use the AI system appropriately by generally knowing how the AI system works and what data it processes, allowing the user to explain the decisions taken by the AI system to the affected person pursuant to Article 68(c).*[10]

And finally, one of the most comprehensive examples of policy guidance on explainable AI has been produced through a collaboration between the Information Commissioner's Office (ICO) and the Alan Turing Institute in the United Kingdom. Explaining decisions made with AI[11] highlights two sub-categories of explanation (process and outcome-based) and six explanation types (rational, responsibility, data, fairness, safety and performance, and consideration of impact) (ICO and Alan Turing Institute 2022).

The Need for Clarity and Context-Specific Guidance

The lack of consistency in how explainability is defined presents a challenge for operationalizing explainable AI policy, especially in different contexts where distinct requirements for explainability are likely to prevail. It only makes sense that considerations such as the purpose, audience and potential impact of an AI implementation should play a role in determining the requirements for explanation, on both technical and human-level grounds. This paper examines key considerations on each of these grounds in the following two sections.

Technical Elements: Post Hoc versus Ante Hoc Explainability

This distinction between models that are inherently explainable on the one hand, and black box models that rely on companion explanatory models on the other, is broadly discussed in the academic literature, but is not well-established in policy. Rarely do policy frameworks or instruments specify what they mean when invoking terms relating to explainability, and even more rarely does the associated guidance parse out the material distinction between ante hoc (that is, inherently explainable models) and post hoc (that is, after-the-fact companion models that estimate how the original task models works) explainability.

[9] *Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* COM(2021) 206 final 2021/0106(COD) art 13, online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.

[10] *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) at amendment 30, online: <www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html>.

[11] See https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/.

This is an unfortunate failure in policy, in large part because the well-documented shortcomings of post hoc techniques render them completely unsuitable for many applications. Whereas there is less reason for concern if a black box model is employed in online advertising or web searches, there is every reason to care about its use in law enforcement, financial services or health care, where an individual's fundamental human rights, health and safety may be impacted. Post hoc explanations should be a course of last resort if they are to be tolerated at all in these use cases, and yet they remain the dominant model class in machine learning.

The prevalence of post hoc explainability techniques is best understood in relation to the popularity of deep learning and neural net algorithms. In 2016, Yoshua Bengio wrote an article celebrating the coming of age of the promise of AI, thanks to the rise of deep learning (Bengio 2016). Deep learning is a sub-form of machine learning that has transformed AI research, powering discovery and ambition in the field. Deep learning methods are often the technique of choice in applications with large data sets, owing to their ease of use and the ability to generate accurate models in certain instances. However, deep learning models are black boxes, and this accuracy of prediction comes at the expense of explainability. Arguably, it is the rise and popularity of deep learning methods that has spawned the movement for explainable AI as stakeholders increasingly recognize the dangers of black box models in high-stakes decision making.

As real as the explainable AI movement is, it is subject to a misguided irony: the widespread and misunderstood use of the term explainable AI. By definition, explainable AI refers specifically to post hoc methods, which are known to suffer from a range of limitations, principally that they generate only an approximation of the functioning of the black box models they are supposedly explaining. In this sense, "explainable AI" would be better known as "estimated AI." And on top of that, post hoc approximations are known to lack fidelity and stability, and their technical shortcomings also mean they are often unable to facilitate a suitable approach to fairness, bias and the absence of discrimination. This is not a trivial shortcoming, with concerns over fairness and the perpetuation of systemic bias at the top of the list of near-term risks from AI. So, in addition to "explainable AI" being more of a "guestimate," there is a very real possibility it is a "bias-perpetuating guestimate" but the lack of genuine explainability makes it difficult to know.

There is a disturbing chasm between the scientific research highlighting the insufficiency of post hoc methods and the volume of explainable AI headlines inferring achievements that cannot actually be claimed today. It has been argued, for example, that the limitations of post hoc methods render them unsuitable as the sole mechanism to guarantee model fairness in high-stakes decision making (Vale, El-Sharif and Ali 2022), and yet we know that black box models (supplemented by post hoc approaches to explainability) are still being applied in high-stakes cases (Rudin 2019) and policy makers are doing little to nothing to change this. It is important that policy makers understand the limitations of post hoc methods to be able to outline guidance, including criteria for when they are not suitable.

From a policy perspective, it is noteworthy that this critical distinction between ante hoc and post hoc explainability is rarely addressed, with a report from the Canadian Office of the Superintendent of Financial Institutions (2023) being a noteworthy exception.[12] Policy instruments, where they exist, tend to be characterized more by an impetus to enact strategic direction for leveraging AI from a research and development perspective and/or implementing explainability from the perspective of de-risking innovation through the reduction of civil rights harms. This leaves a lot of latitude for the providers of explanations, currently complicated by the lack of consistency in policy terminology, standards and implementation procedures.

Foteini Agrafioti, chief science officer at the Royal Bank of Canada, which is recognized as one of the leading banking institutions on AI,[13] spoke to the deficit of explainability in a 2020 interview, specifically addressing the implications for fairness, bias and public trust: "It's important to remember that virtually all of the great machine learning models that are brought to life today through many different products that we use

[12] In the report "Financial Industry Forum on Artificial Intelligence: A Canadian Perspective on Responsible AI," the question of appropriate level of explainability is examined, with inherently interpretable models versus post hoc techniques discussed in the context of criteria such as the party needing explanation and the materiality of the use case.

[13] See https://evidentinsights.com/ai-index/.

today are, unfortunately, unexplainable. You have an input and an output, but you don't really know how the AI got there. For certain sectors — like healthcare and financial services — this is extremely limiting. In areas like lending, which has a serious impact on people's lives, you simply cannot be extending (or not extending) credit without understanding exactly why the algorithm made the decision" (Christensen 2020).

The authors highlight the distinction between post hoc and ante hoc explainability because they believe this distinction is at the core of what matters most about explainability. If explainability is, as they believe, a first among equals among the many principles for evaluating the performance and trustworthiness of AI, then the type and level of that explainability is salient to the overall evaluation task. If explainability really boils down to being able to understand the process and outcome of a model, in a particular context, including the audience receiving the explanation, then how is explainability satisfied with a "guestimate" that potentially holds untruth and bias? This is the terrain of the post hoc explanation.

This is not to suggest that ante hoc explainability is called for in all use cases, but the guidance from policy makers should address the appropriate level and type of explanation according to context, including, specifically, where post hoc explanations are not suitable, or not suitable on their own.

Without a good working definition of explainability, an explicit recognition of the distinction between ante hoc and post hoc techniques and some framing of the implications, the policy guidance that is provided cannot hope to appropriately capture, manage and mitigate the risks and opportunities associated with AI in all contexts. The implication is many references to explainability in policy today are too nebulous to be useful from a technical standpoint.

Human-Level Elements of Explainability

The academic literature on explainable AI is insistent that explainability finds its meaning in the context of the audience for the explanation and not the AI system alone. There are concrete policy examples that make explicit the requirement for meaningful explanation. Canada's Algorithmic Impact Assessment (AIA) tool[14] requires meaningful explanation for AI with assessed impact levels II, III and IV,[15] where these result in the denial of service to a client, or any regulatory action. The United Kingdom's Algorithmic Transparency Recording Standard[16] outlines requirements for public sector bodies using algorithmic tools in decision-making processes that affect members of the public. In both cases, there is a clear purpose to the explanation, an audience, an impact assessment and guidance as to what type of information is appropriate in the explanation. Having said that, neither of these tools get specific as to what may be appropriate for explanation on a technical level, in evaluating what a model is doing to translate inputs into outputs in a high-stakes decision.

Despite some acknowledgement of the role of human-level understanding in explainable AI, there are two areas that need to be better addressed in policy. First, policy should contemplate the wide range of human audiences for these explanations. Whereas policy and regulatory instruments commonly call for human-level understanding, they typically fail to recognize the varying needs and capabilities of diverse human audiences. Even with inherently interpretable models, an explanation may be meaningful and comprehensible to technical personnel and domain experts, but not to other stakeholders. Second, policy *must* link the requirements for explanation to the relevant audience; the needs of the developer troubleshooting a model will inevitably differ from those of the bank customer who is turned down for a loan.

What this discussion really highlights is the role and importance of context in explanation. For explainability to be meaningful, it needs to be appropriately defined in a context, including consideration of factors such as audience, but also other contextual variables such as purpose and impact. This is not always the case in

[14] The Canadian AIA is a component of the Directive on Automated Decision-Making, constituting a tool for AI Impact Assessment. See www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[15] The four levels of the Canadian AIA refer to the level of anticipated impact of the assessed AI, each with its own corresponding set of requirements for things such as peer review, notice and explanation requirement.

[16] See www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies.

policy and the following sections examine, first, what *is happening* with respect to how people conceive of explainability in the governance of models, and then what *should be happening*.

Operationalizing Explainability in Policy

This section moves beyond "background" and builds the case for how to operationalize explainability in policy going forward: using the notion of taxonomy for context-specific guidance; focusing not just on country-level guidance but also on organization-level, where the love of neural nets has led everyone off the quest for true explainability; and the need to think aspirationally and call for true (ante hoc) explainability.

A Context-Based Taxonomy

The activity of operationalizing explainable AI in policy should involve establishing guidance associated with the principle, in a particular context. This may be facilitated through a taxonomy, or a hierarchical classification of questions and contextual variables.

Guidance should start with the legal framework for the use of AI and data including privacy legislation, consumer rights, financial services, competition law, human rights, law enforcement and more. From here the purpose of the explanation, the audience for the explanation and impact assessment of the use case establish the case for guidance. Guidance may include direction on whether a process- or outcome-based explanation is appropriate, as well as the type of explanation (that is, ante hoc versus post hoc) and level of detail required for disclosure.

Various taxonomies have been developed relating to explainable AI and including topics such as the explanation-generating mechanism, the type of explanation, scope of explanation, type of model it can explain and a combination of these features (Markus, Kors and Rijnbeek 2021).

The following provides an example of the categories of questions relating to context and the resultant guidance for policy makers:

- What is explainability (for example, human vs. technical, process vs. outcome-oriented)?
- What is the purpose of explainability (for example, right to explanation, shared understanding, enable user control, right to challenge, model development, evaluation and enhancement, delivery of new insights)?
- Who is the audience for the explainability (for example, developer, user, consumer, subject matter expert, regulator)?
- What is the anticipated risk/impact of the AI system when deployed (for example, low, high)?
- What is the type of explanation required and why?
- What is the specification of explainability requirement; what needs to be shared, what level of detail, in what style, and to whom?

These questions may be conceptualized in a taxonomy as shown in Figure 2.

Jurisdictional versus Model Governance Policy

With some exceptions, current policy guidance on explainable AI emphasizes a jurisdictional (that is, country, state/province or supranational level) perspective, with a spotlight on human-centred considerations such as fundamental rights, and health and economic impacts.

While policy at the jurisdictional level is important, there is also a need for explicit guidance inside enterprises that develop, procure, use and deploy AI. This is in part due to the reality that regulation will vary in breadth and applicability across sectors, use cases and contexts. But even beyond that, if the compliance requirements imposed on institutional-level model governance are mainly limited to the impact-focused guidance provided by jurisdictional regulators, then we collectively miss the opportunity for enterprises to be motivated in the pursuit of explainability methods that serve different but critical purposes including efficiency and innovation.

Industry's enamour for deep learning and performance accuracy has led to a neglect of many of the benefits that can best be conveyed by genuine explainability in the form of ante hoc interpretability, including insights into the

Figure 2: Taxonomy of Explainability Guidance for Policy Makers

Purpose of Explainability

- → Right to explanation
- → Shared understanding
- → Enable user control
- → Right to challenge
- → Model evaluation and enhancement
- → Deliver new insights

Audience of Explainability

Developer	User	Consumer	Regulator

Risk/Impact of AI Use Case

High	Low

Guidance

- → What type of explainability is required?
  - – ante hoc/post hoc, global/local, process/outcome
- → What is the specification of explainability requirement; what needs to be shared, what level of detail, in what style and to whom?

Audience-Specific Guidance
(examples of the different interests and priorities for AI explainability among audiences)

Developers	Users	Consumers	Regulators
→ Understanding model behaviour	→ Establishment of trust and confidence in an AI system	→ Understanding AI product outcomes informs choice	→ Assessment of legal and regulatory compliance
→ Evaluation of model performance	→ Basis for challenge and redress	→ Enables ethical assessment of AI	→ Evaluation of model risks and impacts

Source: Authors.

relative importance of different features (variables) in modelling, the analysis of proxy variables and the demonstration of causal relationships. It is not that these aspects of data science are not being practised; it is that they are being practised poorly, with little ingenuity or innovation.

The academic literature laments the prevalence of post hoc techniques and the erroneous presumption of a tradeoff between performance accuracy and model interpretability (Rudin 2019), but this literature suffers more broadly on two fronts:

- → its examination of the utility of ante hoc explainability and associated methods for the purpose of building and troubleshooting models has not translated into widespread innovation or practical application; and
- → the discussion is characterized by an understandable but short-sighted assumption that there are only two categories of explainable — post hoc and ante hoc or that which is inherently explainable by virtue of being "naturally interpretable."

A New Third Category of Explainable AI

The notion of inherent interpretability may be viewed as the holy grail of explainability and while it is achievable today in certain circumstances, it is constrained by the two main components that go into machine learning models: the type of data and the choice of algorithm. Data scientists refer to a perceived tradeoff between interpretability and predictive performance, owing to the presumption that many types of data do not lend themselves to modelling with anything other than black box algorithms. The trouble is, this alleged tradeoff does not necessarily exist as there are alternative methods, but how are policy makers to know what the technology is capable of when it is deeply technical and subject to nearly constant change through discovery and innovation?

What might innovation look like? As one example, this paper offers the possibility of not just *ante hoc methods* that rely on inherently interpretable algorithms, but novel techniques that facilitate *ex ante feature identification*. The promise of ex ante feature identification is to identify the most important features that relate to a phenomenon of interest, directly from the data and prior to a production model being built. This is exactly the opposite of post hoc explainability, where the black box model comes first, and it is only after the fact that the "guestimate" of how it arrives at its outcome is established. With ex ante feature identification, the features that are employed by a production model are selected *before* it is built, leaving no guessing as to how it generates its outcome. This may not seem like a novel concept, but it is seldom operationalized in the construction of machine learning models, because the prevalent methods do not allow it, especially from data sets that contain a lot of features.

The ex ante identification of "most important features" introduces a game-changing third category of explainable AI method, a true innovation in AI/machine learning where models can be fully interpretable regardless of what algorithm is employed to construct them. In essence, ex ante feature identification is a subclass of ante hoc explainability, where the method is not constrained by the class of the algorithm (that is, one characterized by natural interpretability).

In the absence of technological insight and an awareness of what innovations are taking place, regulatory guidance is likely to be insufficient to drive aspirational requirements, including policy and standards development. Instead, guidance will tend toward the known, tried and true. And downstream, when regulatory guidance flows, pertaining to methods and evaluation of AI explainability, these may be ill-informed or based on partial truths, inadvertently perpetuating bad practices, including an acquiescence toward black box algorithms because they are easy, common and well-accepted.

Policy makers need to confer with technologists at the frontiers of explainable AI to understand what is possible, iterate this in policy and drive the boundaries of best practice. It is through the cross-pollination of ideas that policy innovation will take place. In the case of explainable AI this is particularly true because, ultimately, this is a challenge involving both a technical and a human imperative, where explainability can and should be employed for the purpose of developing *better models* in addition to serving regulatory and compliance requirements relating to human impacts.

The fundamental point that is inadequately captured in policy guidance today is that genuine model explainability — in the form of ex ante model interpretability — leads to models that are better across a range of measures with implications for all the impacts that regulators care about as well as the traditional economic measures that enterprises focus on.

For example, ante hoc methods that enable ex ante feature identification offer the possibility of models that are more parsimonious and more interpretable regardless of what algorithm is used. These same models also lend themselves to unencumbered evaluation across a range of metrics, because the ex ante identification of features means it is always possible to know how the model is translating inputs into outputs. The ability to deliver feature importance insights, opens the door for powerful analytics with value propositions across a wide range of use cases, including a rich analysis of proxy variables.

While it is not the responsibility of policy makers to realize these possibilities, it is important they be considered in the endgame that policy is serving: policy that is designed to protect people, communities and financial systems, to mitigate risk and to stimulate innovation.

For readers who are interested in more technical depth of discussion, there are two appendices containing first, an assessment of different methods or approaches to achieving explainability and second, an example of one method for performing ex ante feature identification, including the possibilities for realizing this third category of explainable AI where models are fully interpretable regardless of the type of data or algorithm.

Recommendations

The development of policy for a more robust approach to explainable AI involves several key considerations that address ethical, technical, legal and practical aspects. Collectively, these considerations help to ensure that AI systems are not only effective but also that they are understandable in technical and human terms. Here are the main considerations:

- → **Definitions:** Policy should distinguish among terms including transparency, explainability and interpretability. These terms address different aspects of how AI systems can be understood, scrutinized and overseen by humans. There needs to be clarity on what they mean to help set objectives, standards and methods for the assessment of AI systems.

- → **Ethical and legal frameworks:** Policy should be grounded in ethical principles and legal frameworks that protect individuals' rights and promote fairness, accountability and non-discrimination. This includes considering the implications of AI decisions on individuals, especially in high-impact use cases and ensuring outcomes are explainable, where the explanation is both meaningful and understandable to the human audience. Explainability should be understood as a requirement both to satisfy evaluation of the full range of ethical and legal principles relating to AI use, and as a necessary ingredient for an appropriate avenue to redress and recourse.

- → **Context-specific guidance:** Policy should provide guidance on explainability requirements for AI that is context dependent, considering the purpose of an explanation, the audience for the explanation and the risk associated with the AI use case. This context-specific guidance should extend to consideration of the technical and human-level requirements for explanation; genuine explainability demands an understanding of how the model arrived at its output and that the explanation is both meaningful and understandable to the appropriate audience.

- → **Post hoc versus ante hoc explanation types:** Policy makers should explicitly consider the suitability of post hoc versus ante hoc approaches to explainability in creating context-dependent guidance, especially for high-impact cases. Recognizing the limitations and potential dangers of post hoc explanations as "guestimates," policy makers should work to outline clear "no-go" zones for their application based on type of use case and impact level.

- → **Use of visualization techniques:** The authors recommend that visualization techniques be applied as much as possible, even when the number of significant features is higher. It is the best tool to gain insight from the data directly, avoiding the approximations that other approaches to model explainability will make. It also allows the understanding of the interaction between the features and what can be expected of the precision of the prediction.

- → **User-centric design guidelines:** Promote the use of user-centric design principles in the design and development of AI systems to help ensure that explanations are understandable to non-technical and non-expert audiences.

- → **Education and literacy:** Promote AI literacy including education initiatives that empower individuals to understand and critically engage with AI systems, making explainability efforts more meaningful and effective.

- → **Guidance should be aspirational:** Policy makers should not be constrained to known or existing methods, but rather should be aspirational in establishing guidance. The impetus for ambitious policy is both to realize practical benefits such as better models and public trust, but also to act as a stimulus for technological innovation — in this case tech for good.

- → **Mandatory explainability standards (high impact):** For high-impact use cases, consider the merits of mandatory explainability standards. This could include regulations pertaining to the development and deployment of AI systems as they are applied to applications with significant implications for individuals' fundamental rights, well-being and safety, such as health care, finance, access to essential services and criminal justice.

- → **Mandatory audit and certification (high impact):** For high-impact use cases, consider the merits of mandatory audit and certification processes. Mandatory audit trails allow for retrospective analysis of AI decisions and outcomes. Mandatory certification facilitates a role for independent third parties for compliance with explainability standards. Both will engender best practices, compliance and trust.

Conclusion

The questions of why policy makers have not been more demanding of genuine explainability in AI systems, and why the distinction between post hoc and ante hoc techniques is so infrequently addressed in policy, were raised in the introduction to this paper. If these issues are partially attributable to policy makers' reticence to aspire to methods beyond the realm of what is known, practised and proven, then it is hoped that by highlighting one ex ante feature identification approach as an example, several things will happen. First is that policy makers will become more educated about the distinct approaches to explainability and the various methods within each class, including newer techniques such as Fiins AI. Second is that policy makers will become more aspirational in establishing guidance, both binding and non-binding.

The pace of change with AI and LLMs is dramatic and faces few challenges. What are the implications for a technology that is forecast to be transformative, with such great impact and so many risks? Do policy makers take their cues from the developers rather than the other way around? Is there so little insistence for explainability in generative AI because it currently seems so unachievable? The authors believe that it is a combination of regulatory impetus, stakeholder interest and competitive pressure that is likely to promote innovation and advancement in the explainability of AI.

The authors find that policy relating to explainability of AI exists but is varied and inconsistent, and this is unfortunate given the special role it plays, both in establishing trust and facilitating high-performance models. Explainability is important because it functions as a first among equals, a requirement for other principles associated with trust and performance.

Extending the two major classes of explainability, the paper presents a third category of explainable AI, one that is interpretable but unconstrained by the need to be hitched to a particular algorithm to enable inherent interpretability. The ability of ex ante feature identification to facilitate causal discovery and the development of fully interpretable models represents an innovation in machine learning, delivering better models, better explanations and greater trust. The paper argues that explainable AI policy needs to be aspirational because the field is moving quickly and should be challenged to catch up, keep up and innovate to meet what are deemed best practices, even if the capabilities do not exist today — or are flying under the radar.

A 2019 article on explainable AI concluded with the lament "let us insist that we do not use black box machine learning models for high-stakes decisions unless no interpretable model can be constructed that achieves the same level of accuracy. It is possible that an interpretable model can always be constructed — we just have not been trying" (Rudin and Radin 2019).

In 2024, let us conclude with the comfort that the capability to always construct an interpretable model does exist today, as illustrated with ex ante feature identification and the Fiins case study. This is not to say the approach is warranted in all use cases, but it should be noted that the call for trustworthy, explainable AI is only one of the reasons to favour such a technique. For many use cases, it is not the explainability of the model, but the value of the ex ante insights that will justify the technique, including a wide-ranging potential set of strategic and operational benefits.

This is a case of regulatory interests converging with the innovation agenda. Policy makers should advance the cause of explainability in a meaningful way, by explicitly calling for the use of ante hoc explainability in high-stakes decisions, but also more generally, in service of competitive advantage. In order to move this forward, policy makers must avoid falling into the trap of taking the easy route, believing in the trade-off between accurate and interpretable AI. They should push barriers and encourage innovation.

Acknowledgements

The authors wish to acknowledge the contributions of colleagues from PolyML, CIGI staff and the second reviewer for very insightful comments.

Appendix 1: Methods and Evaluation of AI Explainability

This section will describe several proposed methods for explainability and some evaluation metrics.

Methods Reviewed

Post hoc explainability — global versus local: As discussed earlier, there are two main paradigms of explainability in machine learning, ante hoc and post hoc. There are two categories of post hoc explainability, global and local. Global explainability refers to how the model will use features to arrive at a prediction. If the model is a black box, and the original training data is not available, it is still possible to get some information. Local explainability refers to how the model performs on a single instance. For example, in credit assessment, local explainability speaks to the ability to understand how a model predicts the probability of loan default in the case of one individual, rather than across an entire population.

Surrogate models: This involves training a simpler, inherently explainable model on the results of the black box model. This will be an approximation of the black box model. For example, a decision tree can be built using sample data, with the target being the result from the black box model. The structure of the tree can be used as an explanation of the black box model. A surrogate model can be global (Molnar 2022) or local as in SLIM (surrogate locally interpretable models) (Hu et al. 2020).

Permutation importance (Breiman 2001): If training data for the model is available this technique can be used to determine feature importance. The model is run with the values of a single feature shuffled. The difference in performance of the model on the permuted data gives an inverse measure of that feature's importance. This is done repeatedly with each feature.

Shapley (Shapley 1953): Evaluate the model on all subsets of feature values in a row. The Shapley value for a feature is the average of all differences between the score for a subset with and without the feature.

SHAP values (Lundberg and Lee 2017): Estimation method for SHapley Additive exPlanations (SHAP) with efficient extensions to specific types of models. SHAP values may be aggregated to give a global measure of feature importance.

Counterfactuals (Mothilal, Sharma and Tan 2020): A counterfactual is a modification of the input values for a case that produces a different prediction from the original case. By progressively modifying input features to determine at what point a prediction changes, it is possible to derive a measure of feature importance for a single case.

Local interpretable model-agnostic explanations or LIME (Ribeiro, Singh and Guestrin 2016): Uses data samples both near and far from a single case to train a simple, inherently explainable model. This is essentially a surrogate model that is only intended to explain the single case and may not necessarily extrapolate.

Gradient-based methods (Sundararajan, Taly and Yan 2017): These methods are used on neural networks where a gradient function is used to associate the inner layers' output to the final output of the neural net.

Ante hoc explainability: Ante hoc explainability is usually taken to mean that a model is created via the use of an algorithm that has natural interpretability, where it is possible to see how inputs have been translated into outputs (for example, a decision tree where each node makes a simple decision, and you can follow the trail of these decisions).

Rashomon (Rudin 2019): This method involves creating many models using different methods. Consider a set of these models that perform roughly equally well and select the model that is most inherently explainable. The success of this will depend on the inherent explainability of the available high-performing methods.

Fiins AI/novel technique for ex ante feature identification (discussed in this paper): Fiins AI (for feature importance insights) is a method that determines a minimal set of features on an ex ante basis, providing the highest predictive power and where models may then be developed on an algorithm-agnostic basis.

Evaluation Metrics

Evaluation metrics include measures of interpretability or human understanding and fidelity or technical accuracy (Markus, Kors and Rijnbeek 2021). Human understanding is a combination of clarity (unambiguous) and parsimony (simple and concise). Technical accuracy is a combination of completeness (explains the entire process) and soundness (correct and truthful).

In Table A.1 these measures refer to the final explanation, not the process for achieving the explanation.

As can be seen, many post hoc methods of explanation do a reasonable job of adding some human understanding to the outcome of a particular prediction. However, the quality of the explanation will always be limited by the quality of the predictor. A post hoc explanation where the prediction was based on many features will still be difficult to understand simply because of its size. And no post hoc process can completely explain how the prediction was made.

Another typical problem is the following: suppose the SHAP values (post hoc) method is used for explainability, pertaining to a prediction from a neural net where there are hundreds or thousands of input features. If there are many correlated and/or similar features, SHAP will report that each of them made a small contribution to the prediction, rather than picking the best representative of the group and identifying a large contribution. We may end up ignoring the most important contribution because it was divided into many features.

It has also been found that the post hoc methods are very unstable. The implications of instability are that if we use slightly different data and/or different models on the same data, the explanations are likely to be completely different. This is mostly due to the problem mentioned above and the fact that a post hoc explanation is an approximation (and in fact may be considered an approximation of an approximation). The post hoc explanation is an approximation of the task model and the task model itself is not necessarily perfect. The lack of stability puts post hoc methods at a serious disadvantage.

From a policy standpoint, these challenges with post hoc methods raise, if not guarantee, the likelihood that guidance will end up focused on the wrong factors. Post hoc methods also frustrate the ability of policy makers to ensure fairness in AI systems and models, as there is no accurate approach to evaluating how inputs are translated into outputs. With black box models, bias on sensitive grounds may go undetected or, alternatively, it may be unjustifiably attributed.

The Rashomon ante hoc method can be said to explain the entire prediction process since an inherently explainable method is chosen, but there are limitations. The capacity of Rashomon to produce inherently explainable methods only holds true if a very small number of best features is used.

As an alternative to the methods reviewed above, including post hoc techniques, approaches that are inherently interpretable but constrained by size and the Rashomon ante hoc method with its own limitations, the paper has presented an ante hoc technique based on ex ante feature identification. Known as Fiins AI for feature importance insights, this technique is algorithm agnostic and, as such, does not suffer from the aforementioned limitations.

The Fiins technique offers a novel approach to mining insights from data that enables the development of inherently interpretable models, even with large and complicated data sets, especially those with many features. The Fiins approach harvests insights directly from data, identifying the most important features in relation to a particular target. It manages to find a very small number of features with high predictive power (or even higher than with all the features). In a situation where there are many correlated features, this approach based on ex ante feature identification will identify the best of them and add the insight that the others may be considered "proxies." This technique can reduce hundreds or thousands of features to an "explainable" number, usually between three and seven, along with their proxies, and ignore those features that are effectively noise.

The Fiins approach provides an explanation based on the data, not on any particular method, so the idea of explaining the model is irrelevant. In this regard, Fiins AI is algorithm-agnostic: it finds the features that are best predictors for whichever model is most suitable. The explainability comes from the very small number of features, typically between three and seven, the proxies of each feature (which helps understand

Table A.1: A Comparison of Explainability Methods

Explainability Method	Ante hoc or Post hoc Global or Local	Human Understanding (low, medium, high)		Technical Accuracy	
		Clarity (Unambiguous)	Parsimony (Simple, concise)	Completeness (Entire process explained)	Soundness (Correct, truthful)
Surrogate	post, local	medium	high	no	no
Permutation importance	post, global	medium	high	no	yes
SHAP values	post, global	high	high	no	no
Shapley	post, global	high	high	no	yes
Counterfactuals	post, local	medium	high	no	no
LIME	post, local	medium	high	no	no
gradient-based	post, global	low	low	no	yes
Rashomon	ante, global	medium	medium	yes	yes
Fiins	ante, global	high	high	n/a	yes

Source: Authors.

the class of the contribution), the relative contribution of each of the features and graphical representations of the interactions of features.

There are two main takeaways from this review. First, despite the continuing popularity of black box models and reliance on post hoc methods for explainability, these methods provide less-than-ideal explanations into how models arrive at their outcomes. Second, even though there are conventional machine learning methods that can be classified as inherently interpretable, these are materially constrained by the limitations of size. The policy implications of this reinforce all the cautions that have been addressed in academic literature, especially pertaining to high-impact decisions.

There needs to be effective correspondence between developers and policy makers, to allow for an informed, iterative and aspirational approach to defining requirements, that balances what is needed in the purest sense, with what is possible today and into the future. In the case of artificial narrow intelligence, policy makers may not have thought much about the distinction between post hoc and ante hoc techniques before reading this paper. And even if they had, they may not have been aware that an innovative technique enabling ante hoc explainability exists.

If a model is built using the Fiins technique or another like it, a model will always be interpretable in that the features the model is using to translate inputs into outputs are readily identifiable. The explanation will be meaningful to any audience if the features themselves are individually and collectively understandable. This level of interpretability is facilitating technically but also in a policy sense, as it provides a clear line of sight through which to evaluate other factors including bias, fairness and the involvement of sensitive variables.

Appendix 2: An Example of Innovation in Explainability

In this section, the paper presents a case study to illustrate the advantages and insights to using ex ante feature identification, using the Fiins method.

Explainable Methods and the Constraints of Size

To understand the value of the innovation to be found in ex ante feature identification, it helps to understand the current state of the art for achieving "inherently interpretable" models. Today, if you want to avoid using a black box, if you want to be confident that you can explain exactly what your model is doing, you will be forced to select from one of a small set of algorithms that possess natural interpretability.

Linear forms, decision trees and nearest neighbours are examples of methods that humans can understand, provided they are reasonably small.

Consider a case where a loan is to be given on the basis of certain criteria that minimize the probability of default. Following analysis, the data is found to fit to a linear form and the predictor might look something like this:

YearlySalary/ 1000 - Age > 20

In this example, it is easy to see that annual salary and individual age are the two criteria that determine whether a loan is to be given or not. For example, a person earning $60,000 per year, who is younger than 40 years of age, will be accepted for a loan. Or a person with a $40,000 salary and 50 years of age will not be accepted. This is easy to understand because linear forms (that is, combinations of features multiplied by numbers) are simple, but also because this linear form is short; if it had 50 or 100 terms (that is, age, salary and many other variables), it would be impossible to grasp.

There is a huge caveat to achieving understandability through the conventionally available methods for inherent interpretability and that is size. These explainable models are only truly understandable when their size is small. No matter how explainable the model is, if it uses 100 features it becomes very difficult, if not impossible, to understand.

These explainable models all fail when they involve real-world sizes. It is an illusion (or a technical urban legend) to think that explainability can be broadly or reliably provided through these types of models in practical settings. This is a highly constrained approach to the achievement of explainability.

Finally, even if the size is moderate and we can partly understand it, we will be understanding what a particular model is doing, not the real problem. The model is likely to be one possible approximation of the prediction problem, not the real situation. In summary, there are too many problems with the currently reigning inherently explainable models to be our tool of choice.

In high-stakes decisions today, one of several things may be happening. In many cases, black box models are still being used, accompanied by post hoc explanations. In other cases, inherently interpretable models may be employed and work well, because even with many features as a starting point, it may be the case that a simple linear form with a small number of features does a good job at prediction and still yields meaningful understandability. Alternatively, there may be cases where for various reasons, including regulatory requirements, an inherently interpretable model is needed, but the constraints of these methods do not suit the problem well, and so in this case it is like trying to fit a square peg into a round hole; valuable insights get missed or mangled.

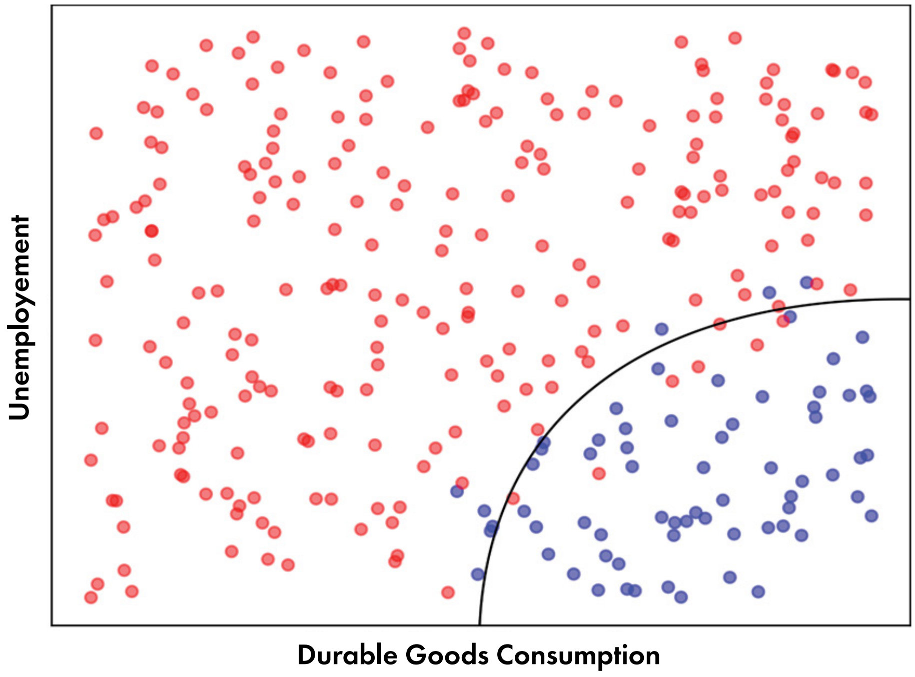
The point is that these decisions are almost entirely left up to developers, with a dearth of appropriate guidance from policy makers at the enterprise or jurisdictional level. Policy makers have both a responsibility and an opportunity to set direction that protects individuals and stimulates innovation.

An Ideal Scenario

In this section, the paper explores the merits of AI models whose explainability is based on a simple structure, thereby allowing visualization to be employed as a powerful tool to facilitate explainability, including meaningful human-level understanding.

First, we will think of a very simple example, one that is not necessarily correct but is useful for

Figure A.1: The Delineation of Features Predicting Credit Card Delinquency — An Example of the Power of Visualization in Explainability



Source: Authors.

illustrative purposes. Suppose that two features, unemployment (U) and durable goods consumption (DG), have 100 percent of the predictive power for credit card delinquency (CCD). Any model using any features will not have a higher predictive power than the best model using U and DG — or together, unemployment and durable goods consumption can perfectly explain credit card delinquency levels.

In this ideal scenario, we can plot the CCD in two dimensions, the U and DG values, and thus be able to visualize how the predictions can be made. For example, we may see that when unemployment is high and durable goods consumption is low, then credit card delinquency is high. With only two features that combine to provide perfect predictive accuracy, this is a simple and clean example to visualize.

An example of this situation is in the graph below, where we see that there are two well-defined areas, consisting of the bottom right corner (that is, in the blue zone, CCD is low) and the rest (that is, in the red zone, CCD is high). This is what we will call the structure of the problem. We can visualize how the features interact to give us an almost perfect prediction of red versus blue. We expect that a good model will be able to use the data properly and make the predictions in an optimal way. But the structure is a property of the data by itself. No model is used to produce the graph. Our goal of explainability is achieved by visualizing the structure. Notice that an extra bonus is the fact that we will have a visual clue on the accuracy of the predictions (that is, how many reds or blues are misplaced).

The ability to use visualization as a tool for facilitating explainability and meaningful understanding can be extended to three features, because it is possible to visualize a structure in three dimensions. It is a bit more challenging to show it on a computer screen, but a good 3D representation added with rotations, transparency and other tricks will allow visualization in 3D. Humans are good at understanding 3D objects, because we are surrounded by them! Three features is a hard boundary: humans cannot visualize objects in higher dimensions.

The bottom line is policy makers should push for approaches that are easily understandable. How a model can be explained leads to meaningful understanding. Methods that allow visualization, with clearly delineated features, are understandable. The precision of the prediction will also be visible, as in the graph above we can see that some points are misplaced. When features can be plotted, they become understandable in how they predict the target. A picture is worth a thousand words.

What happens in real situations? The authors have found that in most real cases using the Fiins method and ex ante feature identification, two or three features explain the lion's share of what is to be explained. It is rare that more than three features are needed to explain more than 90 percent of what is explainable. It may be that up to seven features are needed to have a good model, but the top two or three already give a good picture. Hence by visualizing the structure of the top two or three features we will get a quite complete understanding of the problem.

Contrast this finding and the authors' advocacy for parsimonious, interpretable models with the dominant archetype, the neural net. Neural nets are black box models that are created by an algorithm, directly from data, and even the developers who design them have no real insight into what variables are recruited and combined to make predictions. The models can contain such complicated functions of the variables that any facility to understand falls outside of human capacity.

While it may seem obvious, it should be noted that visualization is a powerful tool that is used in practice by various audiences who need to interpret the outcome of an AI model and when that model can be distilled to a simple structure containing only a few features. In particular, policy makers should push for — and even demand — these types of approaches to explanation in high-stakes decisions.

Works Cited

Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, Francisco Hererra. 2023. "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence." *Information Fusion* 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805.

Balasubramaniam, Nagadivya, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen and Sari Kujala. 2023. "Transparency and explainability of AI systems: From ethical guidelines to requirements." *Information and Software Technology* 159. https://doi.org/10.1016/j.infsof.2023.107197.

Bengio, Yoshua. 2016. "Springtime for AI: The Rise of Deep Learning." *Scientific American*, June 1. www.scientificamerican.com/article/springtime-for-ai-the-rise-of-deep-learning/.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Christensen, Karen. 2020. "RBC's chief science officer talks AI, bias and explainability." Rotman Insights Hub. January. www-2.rotman.utoronto.ca/insightshub/finance-investing-accounting/rbc-chief-science-officer-talk.

Hu, Linwei, Jie Chen, Vijayan N. Nair and Agus Sudjianto. 2020. "Surrogate Locally-Interpretable Models with Supervised Machine Learning Algorithms." Corporate Model Risk, Wells Fargo, July 9. https://doi.org/10.48550/arXiv.2007.14528.

ICO and the Alan Turing Institute 2022. *Explaining decisions made with AI.* Information Commissioners Office and the Alan Turing Institute. October 17. https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence-1-0.pdf.

Islam, Mir Riyanul, Mobyen Uddin Ahmed, Shaibal Barua and Shahina Begum. 2022. "A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Task." *Applied Sciences* 12 (3). https://doi.org/10.3390/app12031353.

Lundberg, Scott and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." 31st Conference on Neural Information Processing Systems, Long Beach, CA. https://arxiv.org/abs/1705.07874.

Markus, Aniek F., Jan A. Kors and Peter R. Rijnbeek. 2021. "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies." *Journal of Biomedical Informatics* 113. https://doi.org/10.1016/j.jbi.2020.103655.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/.

Moradi, Milad and Matthias Samwald. 2021. "Post-hoc explanation of black-box classifiers using confident itemsets." *Expert Systems with Applications* 165. https://doi.org/10.1016/j.eswa.2020.113941.

Mothilal, Ramaravind Kommiya, Amit Sharma and Chenhao Tan. 2020. "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–17. January. https://arxiv.org/abs/1905.07697.

Nannini, Luca, Agathe Balayn and Adam L. Smith. 2023. "Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US and UK." *Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023*, 1198–212. Association for Computing Machinery. https://research.tudelft.nl/en/publications/explainability-in-ai-policies-a-critical-review-of-communications.

NIST. 2021. *Four Principles of Explainable Artificial Intelligence.* NISTIR 8312. https://doi.org/10.6028/NIST.IR.8312.

Office of Science and Technology Policy. 2022. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.* October. www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Office of the Superintendent of Financial Institutions. 2023. "Financial Industry Forum on Artificial Intelligence: A Canadian Perspective on Responsible AI." April 30. www.osfi-bsif.gc.ca/en/about-osfi/reports-publications/financial-industry-forum-artificial-intelligence-canadian-perspective-responsible-ai.

Pantelis, Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23 (1). https://dx.doi.org/10.3390/e23010018.

Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *arXiv*, August 9. https://arxiv.org/pdf/1602.04938.pdf.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 210 (1): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Rudin, Cynthia and Joanna Radin. 2019. "Why Are We Using Black Box Models in AI When We Don't Need to? A Lesson From an Explainable AI Competition." *Harvard Data Science Review* 1 (2). https://doi.org/10.1162/99608f92.5a8a3a3d.

Shapley, L. S. 1953. "A Value for n-Person Games." In *Contributions to the Theory of Games II*, edited by H. Kuhn and A. Tucker, 307–17. Princeton, NJ: Princeton University Press.

Sokol, Kacper and Julie E. Vogt. 2023. "(Un)reasonable Allure of Ante-hoc Interpretability for High-stakes Domains: Transparency Is Necessary but Insufficient for Comprehensibility." *arXiv*, July 10. https://doi.org/10.48550/arXiv.2306.02312.

Sundararajan, Mukund, Ankur Taly and Qiqi Yan. 2017. "Axiomatic Attribution for Deep Networks." *arXiv*, June 13. https://doi.org/10.48550/arXiv.1703.01365.

Vale, Daniel, Ali El-Sharif and Mohammed Ali. 2022. "Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law." *AI and Ethics* 2: 815–26. https://doi.org/10.1007/s43681-022-00142-y.

Centre for International
Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org