AI and the Call for Explainability

Artificial Intelligence (AI) is a term given to a variety of computer applications that automate human cognitive abilities like perception, reasoning, pattern recognition and problem solving. The working definition of AI employed by the government of Canada’s Treasury Board Secretariat is “Information technology that performs tasks that would normally require biological brainpower to accomplish.”¹ AI is what enables many of the autonomous operations that are emerging across various modes of transportation, such as connected and autonomous vehicles, collision avoidance systems, and traffic control measures, making it a key area of importance for Transport Canada. AI also drives many applications that inform wider decision-making related to transportation, featuring in analysis of business and R&D decisions as well.

“Explainability” is the demand and expectation that AI “decisions” be transparent, so that the rationale driving the actions of AI may be understood by those who are affected by those actions. In some basic applications of AI, such as in the business decision-making models that have been employed for many years, demonstrating explainability is relatively straightforward. As AI achieves (and exceeds) human-level accuracy in much more complex tasks and applies emerging techniques such as deep learning, so too does the complexity of the AI decision-making system increase. This in turn makes the ability to understand the exact mechanisms by which AI systems instrumented a course of action more elusive.

In many ways the call for explainability represents a collision between the legal accountability structures put into use prior to the digital age, and the technical limitations to explaining complicated AI processes. As AI applications become increasingly pervasive and AI represents a greater share of our collective decision-making, these demands for accountability and explainability of AI systems will similarly increase. Complicating matters is that there is little guidance on what might constitute an acceptably rigorous “explanation”. In the absence of clarity on what might constitute a satisfactory explanation, accomplishing compliance with undefined “explainability” requirements poses a difficult challenge for regulators and industry alike.

The Explainability Dilemma and the Law

One proposed approach to resolving this “explainability dilemma” has been to call for regulation to limit the functionality of relevant AI technologies to only those applications which are explainable. This approach of course would severely limit the development of the AI sector and limit performance, and this course of action has sparked resistance within the technology community. Canadian AI visionary and Google executive Geoffrey Hinton recently suggested that “regulators (insisting) that you can explain how your AI system works…would be a complete disaster.” Nevertheless, Google released a 2019 white paper demanding government proactivity in setting out standards for explainability. In other words, while researchers and industry seek to avoid a rigid interpretation of explainability, they are also seeking a way they might adhere to this principle in a way that satisfies government.

One of the earliest measures to address the issue of explainability has been the European Union’s (EU) General Data Protection Regulation (GDPR) which is a 2016 EU regulation addressing data protection and privacy. Although based in the EU, the GDPR has emerged as a de facto global standard for many elements of data and AI policy due to the extraterritorial reach of its provisions. GDPR has a wide range of implications for AI and is commonly interpreted as requiring that AI must be explainable. However, a more nuanced perspective demonstrates that there is a great deal of room for interpretation.

The exact relevant text of the legislation, Article 15 specifically, requires that in decisions that are taken by AI (“machine learning” to be precise), relevant parties have the right to access “meaningful information about the logic involved, as well as the significance and envisaged consequences of such processing for the data subject”. This is often surmised as the “right to an explanation” although this is not strictly accurate and a “right to an explanation” does not appear in the text of the GDPR. While the term “right to an explanation” does indeed come from the GDPR, it is not part of the regulation itself but rather appears only in the non-binding recitals section at the end of the document.

Citing the GDPR, many in Canada and abroad have since called for mandatory “explainability” in AI, a development that would significantly impact AI applications and practices. There is however no legal requirement for explainability in Canada, nor is

---

3 N.s. Perspectives on AI Governance, (2019)
4 For more see Rajca, “Right to an Explanation: A Right that Never Was (in GDPR)” March 1st, 2018
there a law which directly addresses explainability. In terms of existing Canadian policy signals, policies, there is documentation from the Treasury Board Secretariat (TBS) which suggests that AI should be “auditable” and that a “meaningful explanation” be provided upon request, but these both fall short of mandating full explainability. Canada’s pending international legal obligations, such as in the United States, Mexico, Canada Agreement (USMCA) and Comprehensive Trans-Pacific Partnership (CTPP), take a different tack and would seem to indicate that a punitive adherence to the principle of AI explainability would be viewed as an undue barrier to trade. By that virtue, these agreements can be viewed to essentially prohibit the punishment of companies for a lack of “explainability” in their AI.

Yet none of these documents represent the final word on explainability since all of the legal signaling about explainability has yet to be tested by the judicial system. This leaves a measure of ambiguity on questions about which (conflicting) signals have supremacy over others. Nor is it clear how the principle of AI explainability would be made to fit within the existing body of Canadian law and policy. On this issue, TBS similarly notes that obligations for explainability are ambiguous and will still largely need to be determined by jurisprudence. TBS has opted to advocate for an audit trail, which while left undefined in the precise measures it requires, embodies several principles including that institutions using AI analyse training data for completeness, follow up to ensure that the system output continues to meets its original intention, and that fully “black box AI” is deployed with care, if at all.

**Addressing Explainability in Ambiguous Circumstances**

With existing legal requirements pertaining to questions of explainability being ambiguous at best- and functionally non-existent in Canadian law at the time of writing- adherence to the calls for explainability need to be addressed cautiously. Any approach to explainability must first and foremost embrace existing legal requirements and principles. Most pertinent to the issue of explainability are these requirements and principles that speak to upholding accountability and establishing clear legal liability. With the fulfilment of these principles driving much of the demand for explainability in the first place, they should be the principal focus of any explainability criteria that are ultimately developed and put into force. Not only will this improve compliance with the existing body of law but would also represent some meaningful progress to a new legal paradigm that is appropriate to AI.

---

7 N.s. “Responsible Artificial Intelligence in the Government of Canada,” April 10th, 2018
8 N.s. “Responsible Artificial Intelligence in the Government of Canada,” April 10th, 2018
Applying existing legal principles of liability and accountability to AI can also help to inject an intelligent sense of proportionality into the conversation about explainability. For instance, in a situation where there is no risk of liability or concerns about establishing a trail of accountability, AI explainability might not in fact be an absolute necessity. In a situation where there are significant liability and accountability risks, the ability to explain a system’s operations becomes decidedly more important. By contrast, unqualified demands for explainability can represent a functionally unattainable standard. Ambiguity about explainability thus creates not only regulatory uncertainty but also represents a risk to private actors.

Perhaps unsurprisingly, industry has joined the chorus of actors seeking more government action in defining the criteria that would satisfy demands for explainability. Many in industry have called for governments to take a more active role in defining the criteria that would be viewed to satisfy the need for an explanation to provide private actors guiding parameters and de-risking future activity. Absent explicit enumeration of these parameters, adherence to the core legal principles underlying calls for explainability can help to iterate real criteria for establishing whether an AI application is sufficiently explainable. On that basis, criteria could eventually be operationalized in a practical way.

Demanding human-level explanations from machines would be problematic since a machine cannot provide the same type of explanation as a human. Instead, explainability requirements must recognize the fundamentals of what each AI technology can reasonably be expected to accomplish in the circumstances where it is being applied. For example, self-driving cars use machine-vision based on convolutional neural networks which power an application that could endanger pedestrians and other motorists. This is rather different than, say, natural language processing which uses a multilayer perceptron (a fairly simple neural network) to mine text for

Box 1: AI Personhood

If a human driver were to hit a pedestrian, legal recourse would be available due to the driver's personal liability. If a vehicle were to hit a pedestrian while driven by AI, the same degree of clear legal recourse is not available because AI is not a legal entity. It has been suggested that this should be rectified by granting AI legal personhood, in the same way as a corporation is a legal person, as a way of assigning liability in such cases.

Yet a group of well-respected thinkers in the realm have firmly rejected this idea on the basis of it being “based on a superficial understanding and overvaluation of the actual capabilities and objectives of even the most advanced AI systems.” In the end, these kinds of determinations about liability will still depend on fault, which will in turn continue to depend on questions blame, responsibility and explanation. “AI Personhood” is not the quick fix it might appear to be.
patterns and trends, and poses no immediate safety risk to anyone as a consequence of its operation.

Each of these situations would surely merit a different standard of explainability based on the risk of malfunction, the capacity to cause harm, and the ability to unpack the process in a way that reveals meaningful information. For instance, an explanation of the frequency and patterns with which certain text appears (natural language processing) is generally lower risk and more readily understandable to a human being than the contours of micro-pixelation in images and how they compare with millions of similar patterns used in various images and sensors (computer vision).

That is to say that a one-size fits all approach to explainability risks papering-over the core technical operations that underlie AI technologies as well as the particularities of the areas to which they are to be applied. A singular standard for compliance with explainability, as one would expect for explanations that come from human beings, is likely to cause at least as many problems as it solves. The variability between AI applications will undoubtedly require thoughtful and nuanced interpretations of how AI systems function with close attention to the context in which they are applied.

### Explainability with Complex AI

Much of the increasing complexity of AI can in many cases be attributed to the growing prevalence of deep neural networks that has taken place in recent years. Compared with simpler AI systems, neural networks and deep learning come with greater challenges to providing an easily intelligible explanation. This is in large part a result of these systems possessing a “hidden layer”, or often multiple hidden layers, which are not readily accessible to interpretation (see Chart 1). In cases with such multilayered neural networks, which represent a novel development, it is important to take a similarly novel approach to an explanation which reflects the capabilities of the technology.

**Chart 1: A Two-Layer Neural Network used in Deep Learning**
One pathway forward is for the explanation to focus less on the operations of the neural network itself but rather to instead focus on the quality and reliability of the data being input into such an AI system. AI systems “learn” from the data to which they have been exposed, and this information is aptly called “training data”. This training data catalyzes adjustments to the neural network and operations of relevant algorithms in the “hidden layer”. The results of the data exposure that forms the training process is ultimately what makes the operations of the AI system challenging to explain since the data has brought about the change, not a designer making intentional adjustments to the hidden layer.

Careful attention to the system inputs—namely the reliability of the training data and the methodology employed—can improve the reliability of the system’s performance across cases. By applying rigorous methodology and carefully controlling information inputs, it is possible to greatly improve the functional control that a system designer has over outputs (see Chart 2). This in turn provides a greater measure of causality for the AI system and therefore a way of interpreting its activities, which itself ultimately constitutes a form of explanation. In other words, instead of finding an explanation for why an AI application opted to behave in a certain way—which may or may not even be possible from a technical standpoint—one could determine what parts of the data or methodology are responsible for the neural network’s development, which in turn explains the system’s behaviour later on.

Chart 2: Potential Points for Probing AI Decision-Making

| Data/Algorithm Inspection | Outcome/Performance Inspection |
This can in turn help ensure that outcomes are replicable and that training datasets are reliable, thereby providing a pathway to a greater measure of explanation. Over time, certain methodologies and data sets will emerge as more reliable than others, reducing the mystery of the AI’s performance across cases. Similarly, having established replicability as a basic principle it becomes conceivable to employ controlled experiments that retroactively identify errors in outputs and validate functional elements. By consequence, it is possible to envision digital “audits” of AI systems that would be able to find errors in data inputs and methodology and, as necessary, replicate the chain of events which led to errors in outputs. Although rather different than the explanation a human being might provide for their decisions, this type of auditing nonetheless similarly provides a pathway to greater of accountability of decisions.

**Addressing the Black Box**

“AI Explainability” is often raised in contrast to the “AI black box”, or the core operational mechanisms of AI which render explanations inaccessible or functionally incomprehensible. While the core AI “decision” may very well lack an obvious or transparent explanation, some have suggested that the issue of the “AI black box” can be circumvented altogether by better design. While AI may naturally gravitate towards a black box model, in many cases it is a relatively straightforward affair to design AI systems in a way that some form of rationalization is readily available. In other words, an appropriate level of explainability can be built-in to the AI system from the outset, rendering an AI decision-making architecture that is conducive to human understanding.

To demonstrate this, first imagine as an example that a black box model is employed to determine individuals who may be likely to have been dishonest in their income tax filings. In this illustration, some undefined data would be entered into an unencumbered AI system (see Chart 3). This system then examines the inputs and concludes that a certain individual should be examined more closely by a tax audit. However, the black box is unable to offer further explanation for why. This kind of unencumbered AI functioning may be commonplace in abstract and experimental uses conducted in labs and universities, but can be troubling as a replacement for human decisions about other humans. Especially so when this situation offers little recourse for appeal in the absence of a justification to begin with.
Chart 3: Unencumbered AI System

Data → Black Box → Output (Decision)

This audit selection system could have instead be designed into modules from the outset (see Chart 4). For instance, one black box system could indicate that the individual in question had many self-reporting inconsistencies in the past, a second notes the individual has history of misfiling, and a third observes an unusual variance in reported income this filing. Taken together, these observations would lead to a decision that the individual should be subject to additional scrutiny. In this example, data was entered into a modulated AI system composed of three black boxes, each of which independently indicated that a certain individual 1) had self-reporting inconsistencies, 2) a history of misfiling, and 3) observed variance in income.

Chart 4: Modulated (“Explainable”) AI System

Data → Black Box 1, Black Box 2, Black Box 3 → Output (Decision)

In this illustration, both of these systems will use AI to conclude that this particular individual may be worthy of further investigation, but in the Modulated AI System, this is done in such a way that intrinsically offers a plausible justification that can be readily understood by human beings. Indeed, a modulated system can provide as valid a justification as could a human being that was relying on a checklist to conduct the same analysis. In this sense, such a system that has been designed for explainability should offer a comparable level of explanation for artificial decisions as is offered up in justification of simple human decisions. Of course, one serious limitation is that while modulation resolves explainability issue for some relatively simple AI systems, this may
not be sufficiently nuanced a solution for the most highly complex AI systems under development, such as autonomous systems.9

With all of this being said, operating on a black-box system may in fact be an advantageous design feature in certain precisely-defined circumstances and can offer several advantages in terms of intellectual property protections and solving collective action problems related to research collaboration. For instance, multiple competing entities could contribute information to train an AI system operating on the basis of a black box, knowing that the system will adopt the insights from their data without presenting the opportunity for this to be reverse-engineered in a way that might divulge proprietary information. This feature can help to propel research forward by reframing the black box as a solution to collective action problems pertaining to intellectual property.

In certain circumstances the natural explainability deficit that comes with AI systems can prove a boon to discovery research. Once again, the key to process explainability and reconciling the black box with existing norms is proportionality. The technical deficits that come with AI and black box systems are morally neutral and should be viewed against the backdrop of the circumstances in which they are employed. Some of the early challenges that have come with AI occurred in where an explanation was desirable, and to be expected in normal circumstances, and where nonetheless a black box model was adopted against the grain of existing norms. As some circumstances exist where full openness is neither required nor desirable, so too are there very precise circumstances where the intrinsic limitations of black box analysis need not be viewed as problematic.

It is also worth considering the degree to which expectations of AI explainability exceed those for human decision-makers at the centre of existing processes. This point was raised by several stakeholders in a round of open public consultations focusing on the primer. Stakeholders noted that explanations given both by AI and human decision-makers are imperfect, but that AI is held to a much higher (and perhaps unrealistic) standard of explainability when compared to what is expected of existing human-centric processes. One particularly salient observation pertained to the limitations of human recall for past decisions as there is a well-documented variance between the reasons for why a particular decision was made at the time, and explanation used to justify that decision after the fact, otherwise known simply as “recall bias” or “response bias”.10

---

9 One suggestion is for AI employing computer vision, including autonomous vehicles, to display a heat map over the optical cues in order to show what which are most significantly contributing to the system’s choices to act at any given time. While this indeed improves a casual observer’s ability to understand and interpret the circumstances influencing the AI system’s navigation, it falls short of total decision-making transparency and may well also prove unsatisfactory for achieving “explainability” as well.

10 For more on recall bias see: Ocal and Babin, 2015.
Stakeholders noted that while AI does come with intrinsic limitations from the black box dilemma, it does not face the limitation of recall or temporal variability that flaw human-centric decisions.

An important question for which there is no consensus at the time of writing is how one might comprehensively compare and contrast the validity and reliability of human and AI decisions. This becomes especially important in circumstances where the adoption of an AI system produces in a net improvement in reliability, transparency and accountability while sacrificing the ability to offer explanations which might be readily understandable to a human observer. This is indeed an important trade-off to consider and stakeholders argued that in present circumstances, the reflex is often to judge AI based principally on its intrinsic limitations compared to human-centric decisions rather than taking a comprehensive view of the net improvement an AI-centric system might offer over existing human centric alternatives. This incongruence with continue to mar the debate and interpretation of explainability into the future.

**Outcome Testing and Performance Inspection**

Current practices for most technologies focus on testing the outcomes of these systems in a variety of circumstances which reflect their likely uses, as well as testing in uncommon circumstances which might expose a vulnerability, known as “edge cases”. The purpose of this testing is to ensure that results of an engineering application reliably fall within an acceptable range of performance. This has long been standard practice for engineering testing of all kinds, including for machine learning and software engineers. This type of testing evaluates outcomes not only on the basis of observed tendencies but also on the fundamentals of scientific knowledge production, which ensures that future designs are based on established and accepted knowledge.

While outcome testing and performance inspection alone have been relied upon in the past for ensuring the safety and reliability of systems, the status quo may not suffice over the long term for applications that are heavily reliant on AI. Much of this is due to questions surrounding the explainability of AI systems’ operations. The degree to which existing procedures can be relied upon in AI systems is a hotly debated topic and one for which scarce consensus has been reached. The purpose here is not to weigh in on the debate in one direction or another, but to make clear some of the parameters of the debate and how they might affect the future of acceptable uses of AI.

Part of the reason for this debate is technical, as potential cases that might confound AI or result in poor performance can be much less intuitive for a prospective human tester then it would be in more standard questions of engineering. For example, in the case of the autonomous Uber vehicle fatality, the AI system was confounded by a bicycle affixed with plastic bags that was being walked across the street instead of being ridden
along the flow of traffic. The system was unable to properly classify this image given a lack of exposure to similar circumstances, given their relative rarity. Indeed, the AI system had been able to reliably classify bicycles in many circumstances, but this odd case resulted in a system failure (see Chart 5). ¹¹

---

**Chart 5: Edge Cases in Car-Based AI (Computer Vision)**

Some edge cases can be reasonably foreseeable and tested for, while others defy human judgement, and others still are nearly impossible to predict.

Source: Koopman, 2019.

The peculiar edge case at play in the Uber fatality has no doubt been corrected for since the accident, but there are likely to be even more peculiar circumstances yet undiscovered, some of which will result in system failures. On one hand these systems will continue to improve as they are able to catalogue an ever greater number of edge cases but on the other hand, increased proliferation of AI systems will also dramatically increase the exposure to potential edge cases. While closed courses and even digital simulations can test edge cases, these approaches are limited to only those situations that can be anticipated in advance by testers, certainly an insufficient criteria for much of the radical innovation that is on track to take place in AI.

This development trajectory and the downstream types of performance testing necessitated will bring many AI applications even further away from a literal interpretation of explainability through performance testing. As system failures inevitably continue to occur throughout the R&D process, whatever explanation exists for justifying performance gaps will be of a very different variety than would have sufficed in pre-AI testing methodology. This is to suggest that the increasingly sophisticated AI applications under development at present will invariably necessitate a highly nuanced, digitally savvy and flexible notion of explainability when it comes to the testing of performance.

Another challenge to status quo performance testing is that compared to physical systems, digital systems tend to evolve at a much more rapid pace. That is to say that digital systems (including AI) do not need to adhere to the long product development cycles that constrain physical systems. As such, new developments like software updates can enter public circulation nearly instantaneously. While the physical design specifications of vehicles, as one example, may only be changed significantly at the annual release of a new model, many AI systems are updated monthly, or even more frequently still. This presents a very practical problem of how traditional outcome or performance testing could be conducted with the needed frequency to keep pace with the development and growing ubiquity of AI.

A digital alternative to performance testing of physical AI systems is a form of digital testing, often referred to as “fuzzing”, which tests edge cases many orders of magnitude higher than could be performed by traditional means by rapidly feeding huge amounts of randomized data into the system in order to expose vulnerabilities. However, given the technical sophistication of the process, “fuzzing” may also prove even more elusive for those seeking a common language explanation. While “fuzzing” may prove a worthwhile alternative to physical testing of AI applications, it does not resolve the explainability dilemma in that it too can elude a common understanding which might satisfy demands for explainability.

Conclusions

With origins in the EU’s GDPR, “AI explainability” has developed into a popular notion, albeit one that is riddled with technical and legal misconceptions. The principle of explainability— that is of making AI applications more transparent and understandable— is a valuable one that should be strived for, but in careful accordance with its technical feasibility. Yet it is important to keep in mind that according to the letter of the law, explainability is neither defined nor required. From a technical standpoint, there are also significant trade-offs that exist between explainability and performance, making it
important that any pursuit of explainability strike a nuanced balance with technological limitations.

Circumventing a potential impasse on the issue of AI explainability will require careful attention. Too rigid adherence to calls for explainability may stifle progress while too rigid adherence to existing legal requirements may generate controversy and fail to protect the public interest. A balance must be struck between letter of the law and the spirit of the law. Uncompromising yet undefined demands for explainability, especially those that are largely separated from pre-existing legal and technical norms about technology, represent a risk both in terms of disruption and in presenting a potential impasse to continued progress.

The priority will be to achieve results that can satisfy the spirit of explainability and can be accomplished by a variety of methods that probe at different parts of AI’s functioning. Employing methods of satisfying demands for transparency and accountability, such as by vetting decision-making inputs, modularizing AI black boxes to provide more information on decisions, and carefully testing performance and robustness, will all contribute to a viable pathway forward. Ultimately, a solution to the explainable AI impasse will likely depend on a combination of approaches that are attuned to the specific AI applications, as well as a clear enumeration of the duties and responsibilities that exist surrounding AI explanations.

Over time, it is conceivable that AI systems will eventually reach a level of standardization and reliability that approaches the kind of global standards that are used in other realms of engineering, and questions of explainability will all but dissipate. A global AI ISO (industrial standards organization) standard is far from being a reality today, but may be an eventual outcome over the long term and across different AI application areas. In the meantime, sophisticated AI applications and the development of processes that are more technically explainable is likely to encumber regulators with new roles and capacity needs in data science and software development.

At the most basic level, emergent circumstances will necessitate growing technological and policy literacy on the part of regulators so they have the capacity to effectively evaluate the operations of these systems. Regulators will also need to develop the capacity to accumulate necessary data to perform digital inspections and AI audits, which may occur through independent data collection, some form of partnership or perhaps even newly emerging tools, such as data trusts. These capacities may be gradually in-sourced over time through improved human resources, or may end up being satisfied through a distributed governance function, such as a partnership or even full outsourcing. While the exact governance mechanisms are not yet clear, the need for enhanced understanding of AI’s implications, such as that of explainability, will be crucial to progress going forward.
Bibliography


