The Impact of Explainable Machine Learning on Innovation in Healthcare: Why Explainability is a Harmful Distraction from Patient Safety Objectives

Tina Dekker
January 21, 2022

1. INTRODUCTION

The growing presence of artificial intelligence (AI) tools in society is driving a policy movement for the responsible development of AI.[1] A principle that is often stated as necessary is explainability,[2] however, the exact meaning of what constitutes explainable AI is left vague. Explainability is generally a concept aimed at elucidating the nature of machine learning (ML) models, which is a subset of AI that learns without explicit programming.[3] Explainable ML has different meanings depending on the context in which it is required.[4] In the healthcare context, some academics are calling for a high degree of ML explainability, *i.e.*, at the model level, to promote patient safety and foster trust in AI.[5] However, requiring explainable ML in healthcare to realize the goals of patient safety and ML accountability is an overstated solution to these goals that fails to consider the cost of explainability to innovation in the field of ML healthcare tools.

---

[1] See *e.g.*, Organization for Economic Co-operation and Development, "OECD AI Principles Overview", online: *OECD* <https://oecd.ai/en/ai-principles>.
[2] See *e.g.*, Organization for Economic Co-operation and Development, "Transparency and explainability (Principle 1.3)", online: *OECD* <https://oecd.ai/en/dashboards/ai-principles/P7>.
[3] Alexander Amini & Ava Soleimany, *MIT 6.S191: Introduction to Deep Learning* (Faculty of Engineering, Massachusetts Institute of Technology, 2019), online: <http://introtodeeplearning.com/2019/index.html>.
[4] See Ankur Teredesai et al, "Explainable Machine Learning Models for Healthcare AI" (26 September 2018) at 18:13, online (video): *Association for Computer Machinery* <https://learning.acm.org/techtalks/healthcareai>; Guang Yang, Quinhao Ye & Jun Xia, "Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond" (2022) 77 Information Fusion 29 at 31.
[5] See *e.g.*, Jocelyn Maclure, "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind" (2021) 31:3 Minds and Machines 421.

"Explainability" as a regulatory requirement or policy objective should be replaced with specific objectives rather than left as a vague requirement open to interpretation; a broad interpretation of explainability often leads to the assumption that ML models must be transparent or, failing that objective, that every decision must be traceable to specific model features or elements. This type of explainability causes more harm than good, both to innovation and patient safety.[6] Developers of healthcare AI tools have a pertinent interest in the intellectual property rights available to them. Generally, where a developer has greater control over its intellectual property rights for its AI, innovation is more likely to flourish in this field.[7] Yet, developers' intellectual property rights will be negatively impacted where model-level explainability is required. It follows that the choice to require ML explainability should be carefully examined to understand how to preserve the economic interests of developers and whether explainability requirements are necessary at all. To understand the latter, a stakeholder approach is necessary to design and use ML in healthcare in a way that reflects stakeholder values, especially the values of providers and patients.

This paper will explore the consequences of explainability to innovation by dissecting the concept of "explainability" into more specific definitions that reflect the true objectives of requiring explainability in the healthcare context. This approach reveals that while a balance between requiring explainable ML and developers' intellectual property interests is possible, the

---

[6] See *e.g.*, Alex John London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability" (2019) 49:1 Hastings Center Report 15; Marzyeh Ghassemi, Luke Oakden-Rayner & Andrew L Beam, "The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care" (2021) 2 Lancet Digital Health 745-750.

[7] World Intellectual Property Organization, "Innovation and Intellectual Property", online: *WIPO* <https://www.wipo.int/ip-outreach/en/ipday/2017/innovation_and_intellectual_property.html>.

reasons for requiring explainability are rebutted by evidence that explainable ML will negatively impact patient safety and provider adoption more than it will promote these objectives.

The remainder of Part 1 will examine the source of explainability as a requirement in ML development and some of it definitional facets. Part 2 will analyze the impact of explainability requirements on developers' incentive to innovate, which is tied to their commercial interest in ML models and the intellectual property protections available to them. Part 3 will discuss the role of explainability in response to safety concerns, demonstrating that explainability is not required to meet objectives of trust and justification, but instead are more likely to hinder the responsible use of AI tools in the healthcare context.

1.1. The Many Facets of Explainability

Much of the discourse about explainable AI has stemmed from the General Data Protection Regulation (GDPR),[8] which provides guidance on European data subjects who are subject to automated decision making. Notably, the GDPR does not provide an explicit right to explanation, although it has been argued that the GDPR indirectly provides for it.[9] Setting aside this debate, the concept of explainable AI now appears frequently in academic discourse about the development of responsible AI, yet the definition of the term itself varies.[10] For example,

[8] EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),* [2016] OJ, L 119/1 [*GDPR*].

[9] The debate about the right to explanation and the GDPR is outside the scope of this work. For more information, see *e.g.*, The Law of Tech, "The Right to Explanation: What's the Debate All About?" (9 May 2021), online: *Medium* <https://thelawoftech.medium.com/the-right-to-an-explanation-whats-the-debate-all-about-2761a45480dc>.

[10] See *e.g.*, Frank Ursin, Cristian Timmermann & Florian Steger, "Explicability of Artificial Intelligence in Radiology: Is a Fifth Bioethical Principle Conceptually Necessary?" (2021) 36:2 Bioethics 143; London, *supra* note 6.

explainable AI can refer to AI that is justifiable, transparent, interpretable, or contestable.[11] More simply, explainability means that an explanation for an ML output is present.[12] The facets of explainability characterize the explanation. For example, a patient who has experienced a clinical harm may seek a *justified* explanation from the provider that explains why or how the provider reached the decision that led to harm. A provider may seek a *transparent* explanation as to how a model weighed different features to produce its output.

Similarly, another facet of explainability is its understandability. Consistent with the notion of contextual explainability, Chowdhury argues that "understandable" AI is most important to users, while "explainable" AI is the concern of engineers and data scientists.[13] This can be interpreted to mean that an *understandable* explanation of ML is one that is targeted to a specific audience, such as the end-user. Another facet of explainability is interpretability, which is often described from a mathematical perspective to a technical audience. Examples of mathematically interpretable models are linear regressions or decision trees.[14] This contrasts with models that are mathematically uninterpretable, namely black box models, that are used for machine learning and deep learning, such as neural networks.[15]

---

[11] Guang Yang, Quinhao Ye & Jun Xia, "Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, Two Showcases and Beyond" (2022) 77 Information Fusion 29 at 31.
[12] This is my definition of explainability; "explainability" is not a recognized word in dictionaries.
[13] Rumman Chowdhury, "Is Explainability Enough? Why We Need Understandable AI" (4 June 2018), online: *Forbes* <https://www.forbes.com/sites/rummanchowdhury/2018/06/04/is-explainability-enough-why-we-need-understandable-ai/?sh=5c28b36c62f4>.
[14] Christopher Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Morrisville, Lulu Press, 2021), Ch 5, online: *GitHub* < https://christophm.github.io/interpretable-ml-book/simple.html>.
[15] *Ibid* at Ch 10.

These definitions are just the tip of the iceberg of how a particular person or group might understand "explainability". Even the narrower facets of explainability are subjective to whomever is defining it; as noted above, an interpretable model has a technical meaning, but the term might be understood differently by an ethicist, for example.

2. THE COST TO INNOVATION IN REQUIRING EXPLAINABILE ML

Developers of healthcare AI tools have a pertinent interest in the intellectual property rights available to them, which will be influenced by the degree of explainability required by regulation for ML use in healthcare. Where the developer has greater control over its intellectual property rights for its ML, innovation is more likely to flourish in this field. Developers will generally fall under one of two categories as being either a third-party developer or a health institution. Both developer types are interested in intellectual property protections for ML healthcare applications that they develop. As will be discussed below, health institutions will likely be the primary developers of such tools.

AI consists of one or more algorithms that are deployed using software. Different types of intellectual property rights are available for software: copyright, trade secrets, and patents. Of these protections, copyright is currently the least useful for AI; copyright does not protect an AI algorithm and only extends to the source code, which is considered a "literary work" within the meaning of the *Copyright Act*.[16] This makes copyright the least desirable intellectual property

[16] *Copyright Act*, RSC, 1985, c C-42, s 3(1). Note that copyright policy for AI is under development, and the scope of copyright protection for AI may change in time. See *e.g.*, Innovation, Science and Economic Development Canada, *A Consultation on a Modern Copyright Framework for Artificial Intelligence and the Internet of Things* (Ottawa:

protection for a developer because copyright does not protect the commercially valuable aspect of AI: the algorithm.[17]

In contrast to copyright, a developer has the most control over trade secrets. A trade secret's inherent value exists in its confidentiality.[18] Trade secrets allow developers to maintain a competitive advantage in the market because they have full control over the disclosure of their product, which can be managed using licensing or confidentiality agreements, or remain entirely secret.[19] Trade secrets can apply to an entire ML algorithm or various elements of it, such as the parameters, weights, or validation information.[20] Moreover, developers have successfully leveraged trade secrets to protect and manage their commercial interest in their AI products.[21] Explainability requirements, which involve some degree of disclosure, thus create a natural tension with trade secrets.

The greatest risks to trade secrets are unauthorized disclosure or reverse engineering of the final product by a third-party. Unauthorized disclosure is managed and enforced through contract law. Reverse engineering, while a concern for certain products, is less of a concern in the context of ML healthcare tools. ML tools intended to improve health outcomes, diagnostics, or logistics all rely on enormous quantities of data, most of which are collected and held by health

---

Innovation, Science and Economic Development Canada, July 2021), online: *Government of Canada* <https://www.ic.gc.ca/eic/site/693.nsf/eng/00316.html>.

[17] Rita Matulionyte, "Reconcililng Trade Secrets and Explainable AI: Face Recognition Technology as a Case Study" (30 November 2021) 44:1 Eur IP Rev 46 (forthcoming) at 1, online: *SSRN* <https://ssrn.com/abstract=3974221>.

[18] Jason Howg, "Unique Trade Secret License Agreement Features" (31 March 2017), online (blog): *BLG* <https://www.blg.com/en/insights/2017/03/unique-trade-secret-license-agreement-features>.

[19] *Ibid*.

[20] Matulionyte, *supra* note 17 at 2.

[21] *Ibid*.

institutions. It follows that many ML models for healthcare are developed by these institutions and licensed to third party developers who may use the models in their business model to, *e.g.*, develop a mobile application.[22] Thus, the most likely competitor is another health institution. Health institutions are generally not in competition with each other, with the mindset of seeking better patient outcomes as the top priority. Therefore, in the health context, the risks associated with trade secrets are minimal. It follows that trade secrets are likely the most desirable intellectual property protection that developers will seek.

Strict regulation regarding explainability at the model level can limit the developer's intellectual property rights and has the potential to create a chilling effect on innovation for ML healthcare tools. Such strict regulation considers explainability in the sense of *transparency* or in the literal sense of knowing the inner workings of the algorithm. In this situation, where public disclosure of how an AI tool works is required, developers are more likely to seek patent protection. Patents involve a *quid pro quo* with the government in which the inventor agrees to publicly disclose their invention and receive exclusive rights over the invention for a period of 20 years in return.[23] However, patenting AI is difficult owing to patent eligibility and disclosure requirements: patents cannot be directed to mathematical concepts or abstract ideas.[24] In addition, as part of the *quid pro quo* of receiving patent rights, the patent description must fully disclose how the invention works.[25] Finally, in Canada, software is not patentable *per se*.[26] In a

[22] See *e.g.*, Hero AI, online: <https://www.heroai.ca/>.
[23] *Patent Act*, RSC, 1985, c P-4, s 44.
[24] *Ibid*, s 27(8).
[25] *Ibid*, s 27(3).
[26] Canadian Intellectual Property Office, *Manual of Patent Office Practice* (Gatineau: Canadian Intellectual Property Office, October 2019, last modified November 2021), Ch 22.

practice notice, the Canadian Intellectual Property Office attempted to clarify the Patent Office's treatment of computer-implemented inventions, noting that an algorithm in itself is not patentable:[27]

> If the computer merely processes the algorithm in a well-known manner and the processing of the algorithm on the computer does not solve any problem in the functioning of the computer, the computer and the algorithm do not form part of a single actual invention that solves a problem related to the manual or productive arts.

The Notice further explains that patentable subject matter of the patent's claims, which legally define the invention, should be tied to a tangible or physical effect.[28] Acquiring a patent for AI—assuming the eligibility and disclosure requirements have been met—requires careful framing of the patent claims such that they are valid. Therefore, requiring explainable AI in the sense of having model transparency creates a commercially unfavorable situation for developers, who must disclose the workings of their AI but cannot protect the inherent commercial value in the algorithm.

A counter argument to this potential paradox for developers is that hospitals and clinics could be subject to strict non-disclosure rules where AI tools are deployed. However, this would quickly become administratively burdensome as the use of AI becomes more commonplace. A second counter argument is that the regulatory approval of AI under the medical device regulatory framework can require model-level disclosure for the purpose of regulatory approval,

---

[27] Canadian Intellectual Property Office, *Patentable Subject Matter under the Patent Act*, Practice Notice (Gatineau: Canadian Intellectual Property Office, last modified 3 November 2020), online: *Government of Canada* <https://www.ic.gc.ca/eic/site/cipointernet-internetopic.nsf/eng/wr04860.html>.
[28] *Ibid*.

while still allowing trade secrets for commercial purposes. This is a valid argument from the intellectual property perspective of incentives to innovate. Indeed, requiring explainability at the Canadian regulatory approval stage is possible, but this returns to the question of why explainability is necessary and whether explainability can meet the objectives it is thought to address.

3. EXPLAINABLE ML IS A POOR RESPONSE TO SAFETY CONCERNS

Explainable ML is most often cited as necessary in healthcare to (1) to foster trust and understanding in patients by offering justification for ML-assisted decisions;[29] and (2) to support accountability through model transparency.[30] Thus, in the context of safety, explainable ML is desired in the sense of having model outputs that are *justified* and *contestable*.

These are important goals, but explainability is more likely to detract from patient safety and reduce ML adoption than promote it.[31] First, even if an ML model is somehow perfectly transparent, the issue transitions into problems of understanding, interpretability, and engagement. Patients who are inundated with complex algorithmic information and data about how an AI tool reached a decision will quickly find that information intractable. The same can be said for providers who are trying to justify a clinical decision that relied on an ML output. This

[29] See *e.g.*, Jocelyn Maclure, "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind" (2021) 31:3 Minds and Machines 421.
[30] A Michael Froomkin, Ian Kerr, & Joelle Pineau, "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning" (2019) 61 Arizona Law Review 33.
[31] See generally the work of Melissa McCradden: Melissa McCradden, "AI Ethics & the Law" (presentation delivered in Michael Da Silva, *Artificial Intelligence in Healthcare* (Faculty of Law, University of Ottawa), 18 January 2022) [unpublished].

suggests that ML tools for healthcare should not be *explainable* at the model level. Instead, these tools should also be *interpretable* and *understandable* to the patients and providers who are using or impacted by the AI tools—at the engagement level.

What constitutes an interpretable or an understandable ML model depends on the context of who is using the ML model, what that person or group is using it for, and how the patient is impacted by the model's output (and to what degree). For example, a deep learning algorithm that demonstrates greater diagnostic accuracy compared to a human professional may be desirable even though how the algorithm achieves its results is unexplainable.[32] In diagnostics, accuracy is the salient clinical consideration and directly comparable to a clinician's diagnostic accuracy without ML.[33] Thus, stakeholders most likely want to know what the accuracy is of the ML model, as well as contextual information about the model, such as the demographic profile of the training data and whether different clinical or demographic data affects the model's accuracy. Both the patient and the clinician would likely need to know, for example, whether a dermatological ML diagnostic tool was trained or tested on people of colour.[34] While such an ML model may not be explainable in the sense of being transparent, the utility and application of the model can be characterized and understood by the clinician who uses it. The clinician can then make an informed choice about how and whether to use the model's output in a specific patient's diagnosis or treatment.

---

[32] London, *supra* note 6 at 16–17.
[33] Melissa D McCradden, "When is Accuracy Off-Target?" (2021) 11:369 Translational Psychiatry 1 at 1.
[34] Sara Gerke, Timo Minssen & I Glenn Cohen, "Ethical and Legal Challenges of Artificially Intelligence-Driven Healthcare" in Adam Bohr & Kaveh Memarzadeh, eds, Artificial Intelligence in Healthcare (Cambridge: Academic Press, 2020) 295 at 304.

Another issue with explainability is that it can detract from a provider's discretion. Jacobs *et al* observed a higher rate of providers who followed incorrect predictions from ML models that included explanations indicating which features influenced the ML output the most.[35] This is a form of automation bias in which providers are overestimating the capabilities of either the ML model itself or the explanation that is tied to it. This problem is exacerbated when the explanations that the ML provides are also clinically irrelevant. Reasons that an ML model (or a secondary explanatory model) provides as an explanation for its output are not necessarily based on information that humans would use to make a similar decision. For example, ML models have been shown to highlight irrelevant regions as informing a diagnosis based on a medical image (*e.g.*, highlighting a shoulder as important in a chest x-ray).[36] In exceptional cases, such information could potentially lead to new medical insights, but for many diagnostics this explanatory information is noise that a provider must sift through and brings into question the clinical accuracy of such models.

Finally, explainable AI is largely premised on the notion that clinical decisions depend on a single empirical output (*e.g.*, a probability or risk score) or binary output (*e.g.*, does the patient likely have cancer: yes or no) from an ML model. For complex clinical decisions, an ML output is merely a single empirical element that informs a decision that also factors in a broader set of elements, including, for example, patient preferences, lifestyle, and financial status.[37] An ML

[35] Maia Jacobs et al, "How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection" (2021) 11:108 Translational Psychiatry 1.
[36] Adriel Saporta et al, "Deep Learning Saliency Maps Do Not Accurately Highlight Diagnostically Relevant Regions for Medical image Interpretation" (2 March 2021), medRxiv, online (DOI): <https://doi.org/10.1101/2021.02.28.21252634>.
[37] McCradden, *supra* note 33 at 1.

output is unlikely to form the sole basis of a complex clinical decision, which makes explainability even less relevant.

From these issues, it becomes clear that a stakeholder approach is necessary to consider how patients, providers, technicians, and other personnel want to engage with ML in the healthcare context. This is particularly important for providers, who must trust the integration of ML into their practice. From the examples above, the answer to increasing engagement and fostering patient safety is not contingent on model-level explainability or even feature/prediction-level explainability. What is more important is that the clinicians, patients, and other stakeholders understand the limitations of a given ML model and the contexts in which the model can be appropriately applied.

4. CONCLUSION

Explainability is an oft-cited requirement in AI policy documents listing principles for the responsible development of AI, and especially ML or black box models. This seems to create a conflict between disclosure, stemming from model level transparency, and the commercial interest that developers have in their ML tools. Where developers are expected to have readily available model-level explanations, their intellectual property protections are limited, which can have a chilling effect on innovation. The best option to promote innovation from an intellectual property perspective is to require disclosure only at the regulatory approval level. However, an understanding of the objectives and effects of explainable ML reveal that context is very important to explainability considerations. Broad requirements for "explainability" without context leaves open many questions from a regulatory perspective that can influence the market

for ML in specific industries. In the healthcare context, explainable ML has many facets, and broad regulatory requirements for explainable ML are likely to cause more harm than good in clinical settings.

This paper has shown that explainable ML is likely to hinder innovation in the field of AI in healthcare while simultaneously failing to meet the safety objectives that some academics argue explainable AI will address. Requiring model level and even prediction level explainable ML is therefore a distraction from pertinent issues regarding the use of AI in healthcare. One such issue is the engagement of providers, patients, and other stakeholders with ML tools. The adoption of ML tools in healthcare relies heavily on the willingness of stakeholders, especially providers, to engage with the tools in the first place. To address such issues, a stakeholder approach should be applied to integrate ML tools into healthcare safely and effectively. A stakeholder approach considers what patients, providers, and others expect from AI when it is used in the clinical setting. These expectations can be operationalized into objectives that address the specific needs of the stakeholders in an equitable manner, such as understanding the appropriate circumstances in when a given ML model's output should be factored into a clinical decision and the weight that factor should receive.