EXPLAINING THE UNEXPLAINABLE:

MEDICAL DECISION-MAKING, AI, AND A RIGHT TO EXPLANATION

Michael Brian Lang

Faculty of Law, McGill University

March 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Laws

Table of Contents

Abstract

Significant decisions in medicine are being increasingly delegated to machines. Automated machine learning models, many of which are thought to be at least as reliable and accurate as human decision-makers, are being used to make decisions about diagnosis, treatment, and care allocation. Though these systems will potentially enhance the quality of health outcomes and contribute to more efficient models of care delivery, they also pose an explanation challenge.

Decisions made by machine learning models are often not accompanied by explanations: it is often technically impossible to know why a machine learning system reaches one decision rather than another. This raises difficult legal and ethical questions about responsibility, equality, and the fundamental principles of procedural law.

This essay explores the degree to which unexplainable decision-making interferes with our conventional ways of understanding the practice and regulation of medicine. I suggest, using medical malpractice as a model, that the challenges posed by unexplainable machine learning may be profound. I describe how, in the face of unexplainable machine learning, several jurisdictions have enacted 'rights to explanation,' including Quebec and the European Union. But these emerging statutory rights are unlikely to respond adequately to the justice implications generated by unexplainable machine learning in medicine. In fact, rights to explanation will probably make things worse.

Résumé

Les décisions importantes en médecine sont de plus en plus déléguées aux machines. Des modèles d'apprentissage machine automatisés, dont beaucoup sont considérés comme au moins aussi fiables et précis que les décideurs humains, sont utilisés pour prendre des décisions en matière de diagnostic, de traitement et de répartition des soins. Bien que ces systèmes soient susceptibles d'améliorer la qualité des résultats en matière de santé et de contribuer à des modèles plus efficaces de prestation de soins, ils posent également un problème d'explication.

Les décisions prises par les modèles d'apprentissage automatique ne sont souvent pas accompagnées d'explications: il est souvent techniquement impossible de savoir pourquoi un système d'apprentissage automatique parvient à une décision plutôt qu'à une autre. Cela soulève des questions juridiques et éthiques difficiles sur la responsabilité, l'égalité et les principes fondamentaux du droit procédural.

Cette thèse explore la mesure dans laquelle la prise de décision inexplicable interfère avec nos façons conventionnelles de comprendre la pratique et la réglementation de la médecine. Je suggère, en utilisant la faute médicale comme modèle, que les défis posés par l'apprentissage automatique inexplicable sont susceptibles d'être profonds. Je décris comment, face à l'apprentissage automatique inexplicable, plusieurs juridictions ont promulgué des 'droits à l'explication,' notamment le Québec et l'Union européenne. Mais il est peu probable que ces droits statutaires émergents répondent de manière adéquate aux implications de justice générées par l'apprentissage automatique inexplicable en médecine. En fait, les droits à l'explication sont susceptibles d'aggraver la situation.

Acknowledgments

I have in the course of preparing this thesis benefitted immensely from the guidance and counsel of my supervisor, Professor Jonas-Sébastien Beaudry and of my co-supervisor, Professor Ma'n H. Zawati, both of whom were extraordinarily giving of their time and insight. This document is miles better than it would have been without their most valuable support.

I am greatly indebted to the Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA) for its generous financial support. I am also deeply grateful for the financial support of McGill University's Research Group on Health and Law and of the Centre de recherche en droit public at the Université de Montréal.

This work was shaped significantly by comments and feedback I received as part of the Ottawa-McGill Graduate Colloquium in Health Law, Policy and Ethics in May 2021.

I have also benefitted enormously from the many conversations I have had over the past two years with my colleagues in the Centre of Genomics and Policy. Their questions and comments in my time as an LL.M. student were both challenging and illuminating. The inimitable Professor Bartha Maria Knoppers was a continual source of inspiration and reading material.

Throughout the drafting of this thesis, Spin kept me from working too hard. Just like our best medical AI, Hannah's incredible patience and boundless support has been unexplainable. I couldn't have done it without her.

INTRODUCTION

Humans are delegating some of our most important decisions to machines. In domains as diverse as finance[1] and forest management,[2] national security[3] and medicine,[4] we are increasingly turning to artificially intelligent (AI) computers to make our decision-making more accurate, more efficient, and under the right conditions, more equitable.[5] AI's promise is perhaps nowhere more acute than in the diagnosis, treatment, and management of disease.[6] In radiology[7] and cardiovascular imaging,[8] AI

[1] See Tom CW Lin, "Artificial Intelligence, Finance, and the Law" (2019) 88 Fordham Law Review 531 at 532 ("The progress and promise presented by artificial intelligence and related new technologies in finance and elsewhere in the economy has been remarkable, though much is yet to be realized. We are just at the beginning of the beginning of the age of artificial intelligence. That said, in just the last few decades alone, we have witnessed significant advances in financial technology made possible in part by artificial intelligence in various aspects of the financial sector").

[2] Sigfredo Fuentes & Eden Jane Tongson, "Implementation of Sensors and Artificial Intelligence for Environmental Hazards Assessment in Urban, Agriculture and Forestry Systems" (2021) 21:6383 Sensors 1 at 1 ("Artificial intelligence (AI), together with robotics, sensors, sensor networks, internet of things (IoT) and machine/deep learning modeling, has reached the forefront towards the goal of increased efficiency in a multitude of application and purpose. The development and application of AI requires specific considerations, approaches, and methodologies. This special issue focused on the applications of AI to environmental systems related to hazard assessment in Urban, Agriculture and Forestry").

[3] Alexander Babuta, Marion Oswald & Ardi Janjeva, "Artificial Intelligence and UK National Security Policy Considerations" (London: Royal United Services Institute for Defence and Security Studies, 2020) at vii ("The research has found that AI offers numerous opportunities for the UK national security community to improve efficiency and effectiveness of existing processes. AI methods can rapidly derive insights from large, disparate datasets and identify connections that would otherwise go unnoticed by human operators. However, in the context of national security and the powers given to UK intelligence agencies, use of AI could give rise to additional privacy and human rights considerations which would need to be assessed within the existing legal and regulatory framework").

[4] Eric Topol, "High-performance medicine: the convergence of human and artificial intelligence" (2019) 25 Nature Medicine 44 at 44 ("Almost every type of clinician, ranging from specialty doctor to paramedic, will be using AI technology, and in particular deep learning, in the future").

[5] See e.g. Arthur Rizer & Caleb Watney, "Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just" (2018) 23 Texas Review of Law & Politics 181 at 195 ("With more accurate forecasts regarding the small share of defendants who actually pose a significantly higher risk of skipping trial or being a danger to the community, application of this algorithm could potentially cut crime, cut the jail population, save taxpayers money, and produce a more equitable system of justice").

[6] See Pearse A Keane & Eric J Topol, "With an eye to AI and autonomous diagnosis" (2018) 1:40 NPJ Digital Medicine 1 at 1 ("Given the potentially transformative potential of AI for healthcare (in particular a technique referred to as 'deep learning')—but also its associated hype—this lays an important foundation for future translation of such technologies to routine clinical practice").

[7] Oleg S Pianykh et al, "Continuous Learning AI in Radiology: Implementation Principles and Early Applications" (2020) 297:1 Radiology 6 at 6 ("During the past few years, the radiology community has seen a rapid rise in the potential of artificial intelligence (AI) to transform many radiology applications, from medical image interpretation to clinical and operational decision making").

[8] Maxime Sermesant et al, "Applications of artificial intelligence in cardiovascular imaging" (2021) 18 Nature Reviews Cardiology 600 at 600 ("Cardiovascular imaging is one of the most active clinical applications of AI because of the

models have exhibited a remarkable capacity to interpret images, identify patterns, and categorize large datasets.[9] In certain applications, AI is likely to soon consistently outperform human clinicians on measures of diagnostic accuracy.[10] This likely means that AI's adoption in medicine will on the whole contribute to faster and more accurate diagnosis, more precisely targeted therapy, and better patient outcomes.[11] But medical AI might also be uniquely complicated.

One source of considerable consternation for lawyers and policymakers is that some of our best AI is technically unexplainable. Models that apply machine learning methods to identify patterns, understand natural language, and interpret medical images, are often so architecturally complex that even the humans who initially programmed them have no clear idea how the models work.[12] This greatly frustrates our ability to conduct effective *ex post* review of a model's operations and outputs. In the medical context, unclear explanation might create substantial uncertainty about how clinical professionals and regulators should engage with these kinds of systems. Without knowing precisely how a medical AI model operates, for example, it might be difficult to determine the conditions under

challenges associated with processing images of a beating organ. AI has been applied to all medical imaging modalities, from 2D and 3D images to temporal sequences8 derived from cardiac MRI, CT, nuclear imaging or ultrasonography").

[9] Hayit Greenspan et al, "Deep learning in medical imaging: overview and future promise of an exciting new technique" (2016) 35 IEEE Transactions on Medical Imaging 1153 at 1153 ("In particular, convolutional neural networks (CNNs) have proven to be powerful tools for a broad range of computer vision tasks. Deep CNNs automatically learn mid-level and high-level abstractions obtained from raw data (e.g., images). Recent results indicate that the generic descriptors extracted from CNNs are extremely effective in object recognition and localization in natural images").

[10] David Killock, "AI outperforms radiologists in mammographic screening" (2020) 17 Nature Reviews Clinical Oncology 134 at 134 ("When used to provide a rapid second opinion as part of the double-reading process used in the UK, the accuracy of the AI system was non-inferior to serial reading by two radiologists, and the simulated workload of the second reader was reduced by 88%").

[11] See e.g. Susanne Gaube et al, "Do as AI say: susceptibility in deployment of clinical decision-aids" (2021) 31 NPJ Digital Medicine 1 at 1 ("AI systems will only be able to provide real clinical benefit if the physicians using them are able to balance trust and skepticism. If physicians do not trust the technology, they will not use it, but blind trust in the technology can lead to medical errors").

[12] See W Nicholson Price III, "Regulating Black Box Medicine" (2017) 116 Michigan Law Review 421 at 429–430 ("Black-box medicine is 'the use of opaque computational models to make decisions related to health care' and is the focus of the remainder of this Article. It is the subset of algorithmic medicine where the algorithms are unavoidably opaque, whether those algorithms are used in an mHealth context or in other systems. Typically, such algorithms are derived from large datasets of health information using sophisticated machine-learning techniques and reflect complex underlying biological relationships").

which the model's use in patient care is likely to be clinically justifiable. This is a particular problem from the perspective of medical professional regulation, for both medical colleges and malpractice law generally expect clinicians to conduct themselves according to the standard of the reasonable professional. But knowing what is reasonable in the use of an unexplainable model might turn out to be remarkably difficult. Inasmuch as a model's internal functions are impossible to scrutinize in detail, their causal effects on patients might be difficult to monitor and review. Jurists might find it particularly challenging to coherently allocate responsibility for injuries arising from the use of an effectively inscrutable medical implement.

These problems are at once conceptual and practical. They are conceptual insofar as the unexplainable nature of certain medical AI models appears to be in tension with the foundational logic of medical malpractice law, a logic that depends in no small measure on the availability of clearly delineated standards of practice and coherent factual accounts of fault. These problems are practical insofar as worries about professional liability, whether borne out in case law or not, could have the effect of reducing the clinical uptake of medical AI.[13] To the degree we think medical AI will operate to produce better clinical outcomes, professional reticence to use unexplainable models might have the effect of depriving unwell patients of better alternative treatment. In parallel with conceptual limitations on the operation of malpractice law, unexplainable AI may have the effect of frustrating the capacity of injured persons to make out a compelling case for redress. If the unexplainable nature of medical AI decision-making prompts confusion about one or more of the essential elements of a malpractice claim, then this could have the practical effect of leaving injured persons without recourse. Lawyers and policymakers have increasingly been expressing serious

[13] Kevin Tobia, Aileen Nielsen & Alexander Stremitzer, "When Does Physician Use of AI Increase Liability?" (2021) 62:1 Journal of Nuclear Medicine 17 at 17 ("Legal scholars have cautioned that tort law may create a substantial legal barrier to physicians' uptake of AI recommendations: accepting certain AI recommendations may increase physicians' risk of liability in medical malpractice. In particular, given tort law's privileging of standard care, physicians who accept a personalized AI recommendation to provide nonstandard care would increase their risk of medical malpractice liability").

concern that unexplainable AI operates in disjunction with a whole range of established legal regimes and practices.[14] One potentially attractive solution would have the law combat unclarity about AI directly, to require explanation whenever a functionally complex or inscrutable model is used for decision-making that affects individual interests. This is effectively what rights to explanation propose, that decisions produced by otherwise unexplainable AI be accompanied by an accounting of the reasons and factors for which the decision was made. Rights of this variety have been a favourite tool for the regulation of AI, appearing equally in international ethics norms[15] as in statutory law. [16] But there is much uncertainty about how rights to explanation will operate in practice and whether they will interact coherently with existing legal frameworks.[17] This essay considers rights to explanation in medicine. I aim to address two specific questions: (1) whether unexplainable medical

[14] Hannah R Sullivan & Scott J Schweikart, "Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?" (2019) 21:2 American Medical Association Journal of Ethics 160 at 164 ("The rise of black-box AI and its use in medicine complicates application of existing tort law when trying to resolve claims of malpractice. If a patient becomes injured by use of an AI technology (black-box AI in particular), current legal models are insufficient to address the realities of these innovations. New legal solutions that craft novel legal standards and models that address the nature of AI, such as AI personhood or common enterprise liability, are necessary to have a fair and predictable legal doctrine for AI- related medical malpractice"); Yavar Bathaee, "The Artificial Intelligence Black Box and The Failure of Intent and Causation" 31:2 (2018) Harvard Journal of Law & Technology 889 at 891 ("The reason intent and causation may fail to function is because of the nature of the machine-learning algorithms on which modern AI are commonly built. These algorithms are capable of learning from massive amounts of data, and once that data is internalized, they are capable of making decisions experientially or intuitively like humans. This means that for the first time, computers are no longer merely executing detailed pre-written instructions but are capable of arriving at dynamic solutions to problems based on patterns in data that hu- mans may not even be able to perceive. This new approach comes at a price, however, as many of these algorithms can be black boxes, even to their creators").

[15] OECD, Council on Artificial Intelligence, *Recommendation of the Council on Artificial Intelligence* C(2019)34, C/MIN(2019)3/FINAL ("AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: i.to foster a general understanding of AI systems, ii.to make stakeholders aware of their interactions with AI systems, including in the workplace, iii.to enable those affected by an AI system to understand the outcome, and, iv.to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision").

[16] See *An Act to modernize legislative provisions as regards the protection of personal information*, CQLR, c 25, s 102 ("Any person carrying on an enterprise who uses personal information to render a decision based exclusively on an automated processing of such information must, at the time of or before the decision, inform the person concerned accordingly. He must also inform the person concerned, at the latter's request, (2) of the reasons and the principal factors and parameters that led to the decision") [Bill 64].

[17] See e.g. Sandra Wachter, Brent Mittelstadt & Chris Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" (2018) 31:2 Harvard Journal of Law & Technology 841 at 842 ("There has been much discussion of the existence of a 'right to explanation' in the EU General Data Protection Regulation ('GDPR'), and its merits and disadvantages.' Attempts to implement a right to explanation that opens the 'black box' to provide insight into the internal decision-making process of algorithms face four major legal and technical barriers").

AI is likely to complicate the practice and regulation of medicine as interpreted through the lens of medical malpractice law and (2) whether challenges for medical practice and regulation are likely to be remedied by a statutory right to explanation. I answer the first of these questions in the affirmative and the second in the negative. While unexplainable AI creates uncertainty for medicine, the right to explanation fails as a response and almost certainly makes things worse. While the lawyer's reflex might be to explain that which is unfamiliar or uncertain, legislating explanation for medical AI will not work. I argue that explanation is illusory, more likely to muddy the waters than clear them.

Methods & approach

This essay consists of three chapters. In the first, I introduce foundational concepts in artificial intelligence. I survey how AI might be defined and I conceptualize how certain AI models may be technically unexplainable. I suggest that certain unexplainable models are *deeply* unexplainable, for which not even the model's initial programmer will be able to understand how the program works. I identify some of the models approved by Health Canada that may be deeply unexplainable. I do this by performing an iterative search of Health Canada's Medical Devices Active Licence Listing, a reference tool documenting basic information about medical devices that have been approved by Health Canada's Medical Devices Bureau,[18] and cross-referencing approved devices against those approved by the Food & Drug Administration in the United States.[19] I will provide more precise methodological detail below. In the second chapter, I argue that unexplainable AI has significant implications for the practice and regulation of medicine. I do this by considering how the

[18] Health Canada, "Medical Devices Active Licences Search" (2021), online: <https://health-products.canada.ca/mdall-limh/prepareSearch-preparerRecherche.do?type=active> ("MDALL contains product-specific information on all medical devices that are currently licensed for sale in Canada, or have been licensed in the past") [Active Licences Search].

[19] See Stan Benjamens, Pranavsingh Dhunnoo & Bertalan Meskó, "The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database" (2020) 3 Nature Digital Medicine 1 at 2 ("The purpose of this paper therefore was threefold: (1) to provide an insight into the currently available AI/ML-based medical devices and algorithms that have been approved by the FDA; (2) to create an up-to-date database of FDA approvals in this field that welcomes submissions and might serve as the database that the FDA should have; and (3) to raise awareness of the importance of regulatory bodies clearly stating whether a medical device is AI/ML based").

unexplainable character of certain automated decision-making processes is likely to complicate the law of medical malpractice. This work combines two general methodological approaches. On the one hand, I engage in exposition of the law of medical malpractice as it is presently constituted in Quebec and common law Canada.[20] I outline, drawing on case law and commentary, how doctrines surrounding consent, fault, and the causation of injury are framed and understood by courts and lawyers. On the other hand, I engage in a kind of prospective doctrinal analysis, drawing on law and policy scholarship to imagine how the adoption of unexplainable AI might interfere with the law as it is presently imagined.[21] I argue that unexplainable AI interferes significantly with how the law presently thinks about medical practice. In the third chapter, I consider whether the right to explanation might address challenges for medical practice and regulation raised by unexplainable AI. I employ an expository and comparative method to describe how rights to explanation are constructed in the European Union's GDPR, in Quebec's Bill 64, and in proposed reform to Canada's *Personal Information Protection and Electronic Documents Act*. I describe what rights to explanation require in respect of the use of automated decision-making and consider how it might apply in the medical context. I draw on technical and policy scholarship to argue that rights to explanation do not remedy the challenges raised by unexplainable AI and probably end up making things worse.

As a preliminary note, I do not in this essay resolve the many regulatory challenges generated by unexplainable AI with any kind of finality. I generally leave open, for example, the question of how automated decision-making should be used in medicine. My primary purpose here is to suggest

[20] See e.g. Paul Chynoweth, "Legal Research" in Andrew Knight & Leslie Ruddock, eds, *Advanced research methods in the built environment* (Chichester: Wiley-Blackwell, 2008) at 31 ("The applied form of doctrinal research is concerned with the systematic presentation and explanation of particular legal doctrines and is therefore referred to as the 'expository' tradition in legal research. This form of scholarship has always been the dominant form of academic legal research) and has an important role to play in the development of legal doctrines through the publication of conventional legal treatises, articles and textbooks").

[21] See Terry Hutchinson, "The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law" (2015) 3 Erasmus Law Review 130 at 131 ("Nevertheless, legal academic success has been measured within a doctrinal methodology framework, which includes the tracing of legal precedent and legislative interpretation. The essential features of doctrinal scholarship involve 'a critical conceptual analysis of all relevant legislation and case law to reveal a statement of the law relevant to the matter under investigation'").

that the challenges warrant careful attention and that recently popular proposals to regulate explanation are both unproductive and likely harmful. In the place of explanation, it may be fit for regulators to approach unexplainable AI as a not wholly unique phenomenon. Our institutions and concepts may well be capable of moderating AI's risks without the interference of conceptually imprudent attempts to force explanation. To be sure, much of what I say in this essay may apply to varying degrees in contexts other than medicine. While the medical context reveals some of the most serious problems for unexplainable AI, it also potentially points its way toward a solution. Though it might naturally make us uncomfortable to depend on unexplainable programs, we in fact do so all the time. We routinely entrust our health to processes that are, if not unexplainable, at least unexplained. It may be worth thinking about unexplainable AI, I suggest, on similar terms.

CHAPTER ONE: AI and UNEXPLAINABLE DECISION-MAKING in MEDICINE

This first chapter summarizes the adoption of unexplainable artificial intelligence in medicine. It does this in three parts. First, I give an overview of competing conceptions artificial intelligence. I summarize the field's history and suggest that a human-defined process approach to defining AI should be preferred to an approach centered on a human-machine analogy. Second, I describe how certain AI systems, notably deep learning algorithms, are technically unexplainable. I explain what this means and briefly foreground how technically unexplainable models may pose challenges for the law. Third, I suggest that deep learning's unexplainable character might have special resonance in healthcare. I consider some of the AI systems used in Canadian healthcare and document their basic functions, suggesting they may play significant a role disrupting modern medical practice.

1. *'Defining' Artificial Intelligence*

Alan Turing in his monumental 1950 paper "Computing Machinery and Intelligence" describes a fanciful idea: that machines might one day be able to think.[22] In some distant future, Turing says, machines will compete with humans in every intellectual field.[23] He envisions computers capable of engaging in the kind of abstract thinking required to play complex strategy games, identify patterns, or solve manually unsolvable mathematical equations. Though the computers of Turing's era could not plausibly have been said to 'think,' contemporary models can perform an impressive diversity of tasks that were once the sole domain of human beings. Only three decades after Turing's early death,

[22] See AM Turing, "Computing Machinery and Intelligence" (1950) 59:236 Mind 433 ("We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'").

[23] Turing, *ibid* at 460 ("We may hope that machines will eventually compete with men in all purely intellectual fields").

the best human chess players in the world were being beaten by computers.[24] Since then, artificial intelligence (AI) has proliferated exceptionally rapidly, extending its influence in fields as diverse as agriculture,[25] infrastructure,[26] finance,[27] and the practice of law.[28] It is likely that few areas of human life will be spared AI's impact.[29] AI models promise to do much of what human beings can do, only more efficiently, accurately, and cheaply. That a significant number of human workers could one day be replaced by machines is not merely a notion drawn from science fiction, but a real and insistent concern for consultants and policymakers.[30] But despite impressive technical and philosophical advancement in our understanding of computerized intelligence, AI remains surprisingly difficult to concisely and persuasively define. Little more can be said now than in Turing's era about what AI *is*.[31] This first part of the chapter gives an overview of two possible ways of defining AI: by analogy with human intelligence, or in terms of certain human-defined processes. While both are prominent candidate approaches to understanding AI's scope, I ultimately suspect that thinking about AI in terms

[24] See DNL Levy, "How Will Chess Programs Beat Kasparov?" in TA Marsland & J Schaeffer, *Computers, Chess, and Cognition* (New York: Springer, 1990) 47–52 at 47. See also Demis Hassabis, "Artificial Intelligence: Chess match of the century" (2017) 544 Nature 413.

[25] See Michael J Smith, "Getting value from artificial intelligence in agriculture" (2020) 60 Animal Production Science 46.

[26] See e.g. Lingling Tian, Juncheng Jiang & L Tian, "Safety analysis of traffic flow characteristics of highway tunnel based on artificial intelligence flow net algorithm" (2019) 22 Cluster Computing 573.

[27] See Tom CW Lin, "Artificial Intelligence, Finance, and the Law" (2019) 88 Fordham LJ 531.

[28] See Simon Stern, "Artificial Intelligence, Technology, and the Law" (2019) 68 UTLJ 1.

[29] See Joanna J Bryson, "The future of AI's impact on society" (2019) MIT Technology Review.

[30] See Michael Chui, James Manyika & Mehdi Miremadi, "Where machines could replace humans—and where they can't (yet)" (2016) McKinsey Quarterly; Lori G Kletzer, "The Question with AI Isn't Whether We'll Lose Our Jobs — It's How Much We'll Get Paid" (2018) Harvard Business Review; Mohammad Hossein Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making" (2018) 61:4 Business Horizons 577.

[31] See Selmer Bringsjord & Naveen Sundar Govindarajulu, "Artificial Intelligence" in Edward N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2020) at s 2 ("So far we have been proceeding as if we have a firm and precise grasp of the nature of AI. But what exactly is AI? Philosophers arguably know better than anyone that precisely defining a particular discipline to the satisfaction of all relevant parties (including those working in the discipline itself) can be acutely challenging. Philosophers of science certainly have proposed credible accounts of what constitutes at least the general shape and texture of a given field of science and/or engineering, but what exactly is the agreed-upon definition of physics? What about biology? What, for that matter, is philosophy, exactly? These are remarkably difficult, maybe even eternally unanswerable, questions, especially if the target is a consensus definition").

of human-defined processes provides a firmer foundation for thinking about the kinds of phenomena of interest in the field.

a. Human-Machine Analogy

It might be that defining AI will be unavoidably difficult, for many of our best candidate definitions depend on a potentially dubious analogy between human and computerized intelligence.[32] There are two problems with an approach to defining AI that searches for signs of human intelligence in machines. The first is that we understand quite little about human cognition. Foundational questions about what it means to be conscious, how brain physiology relates to intelligence, and whether human cognition resembles cognition in other creatures, are largely unanswered.[33] Significant debate persists in cognitive psychology and in other fields about the essential nature of human intelligence.[34] To the extent our definitions of AI rely on an analogy with an unsettled and contentious construct, they might be inescapably unsatisfying. The second problem is that, even if we understood a great deal more about human intelligence such that an analogy between it and computerized intelligence could be made, it is unclear why we should think that these phenomena are fundamentally comparable. On the surface, human beings are entities that are very different than computers. Even if most humans do not intimately understand the physiological mechanics of human cognition, we typically have a reasonable sense of how thought looks and feels. We can identify evidence of thinking in many of

[32] John McCarthy, "What is Artificial Intelligence" (2004) Stanford University 2 at 2, online: <https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf>.

[33] See e.g. David Chalmers, "Facing Up to the Problem of Consciousness" (1995) 2:3 Journal of Consciousness Studies 200 (Chalmers argues that consciousness is an ambiguous concept, referring to a large number of different phenomena. He distinguishes 'easy' problems of consciousness from 'hard' problems, or those that resist the usual methods of cognitive science. Probably the most significant hard problem is that of experience, or of what it is like to be a conscious organism. Though we know with certainty that this phenomenon exists and that many kinds of organisms have conscious experiences, it is very difficult to explain how such experiences work or why they exist. Though AI systems are probably not subjects of conscious experience, the hard problem underscores just how little we know about everyday cognitive phenomena. Knowing what it means to think, in other words, is probably not straightforward).

[34] See e.g. Kirsten Hilger, Makoto Fukushima, Olaf Sporns & Christian J Fiebach, "Temporal stability of functional brain modules associated with human intelligence" (2019) 41:2 Human Brain Mapping 362.

our fellow organisms, even those with whom we do not share a common language or whose behaviours and practices are very different than our own. It scarcely needs saying that modern computers do not typically give the same kind of cognitive feedback as conscious, living, breathing, naturally occurring animals. Even the most sophisticated machines are effectively electronic boxes, composed of welded metal and silicon chips. It is intuitively very unusual, perhaps even frightening, to ascribe thought to such entities.[35]

This basic analogy between human and machine intelligence is nevertheless at the heart of many conventional definitions of AI. Turing's influential view, one of the earliest significant candidate conceptions of AI, is that we can properly attribute intelligence to a computer capable of passing an imitation test. The imitation test involves a neutral human observer communicating with two anonymous interlocutors. One of the anonymous interlocutors is human, while the other is a machine. We can reasonably ascribe intelligence to a computer, Turing argues, where the neutral observer would be unable to determine confidently which of their interlocutors was the machine.[36] Some interpreters reformulate Turing's imitation test such that it focuses more directly on the capacity of a machine to "communicate in natural language in a manner indistinguishable from that of a human

[35] See e.g. Appa Rao Korukonda, "Taking stock of the Turing test: a review, analysis, and appraisal of issues surrounding thinking machines" (2003) 58 International Journal of Human-Computer Studies 240 at 242–243 (the 'heads-in-sand' objection to AI in effect argues that human beings are special kinds of creatures. Much of what makes us special is the capacity to think. A worldview that allows for thinking machines displaces our unique character).

[36] See Turing, *supra* note 22 at 433 (Turing describes the imitation game in the following way: "It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus: C: Will X please tell me the length of his or her hair?... We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'" Importantly, though Turing's initiation formulation of the imitation test turns on the capacity of a machine to convincingly mislead a human observer on a question of the former's gender, it is not clear that there is anything strictly significant about the gender assessing abilities of the machine. It could be, Turing appears to allow, that the machine could be engaged in any sort of convincing human imitation. See James H Moor, "The Status and Future of the Turing Test" 11 (2001) Minds & Machines 77 at 78–79.).

being."[37] More contemporary conceptions of AI adopt a similar strategy. One relatively recent textbook suggests thinking about AI as a field broadly interested in "creating machines which solve problems in a way which, done by humans, require intelligence."[38] Additional conceptions refer to AI variously as a program "which in an arbitrary world will cope no worse than a human,"[39] or that is capable of achieving "human or even superhuman levels of performance across a variety of tasks."[40] As AI is increasingly the subject of formal state regulation, legislators and government agencies have also established their own definitions of AI. These also sometimes depend on the human-computer intelligence analogy. Canada's Treasury Board Secretariat, for example, which administers the operation of the federal bureaucracy, describes AI as information technology "that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems."[41] This definition, expressed in a 2019 "Directive on Automated Decision-Making," is intended to delimit the adoption of AI systems by federal institutions, to reduce risks for government entities and Canadians alike, and to produce "efficient,

[37] Ayse Pinar Saygin, Ilyas Cicekli & Varol Akman, "Turing Test: 50 Years Later" (2000) 10 Minds & Machines 463 at 468 ("It is generally agreed that the gender issue and the number of participants are not to be followed strictly in attempts to pass, criticize or defend the TT. Even Turing himself, in the subsequent sections of 'Computing Machinery and Intelligence', sometimes ignores these issues and focuses on the question: "Can machines communicate in natural language in a manner indistinguishable from that of a human being?").

[38] Crina Grosan & Ajith Abraham, *Intelligent Systems: A Modern Approach* (New York: Springer Publishing, 2011) at 1.

[39] Dimiter Dobrev, "Formal Definition of Artificial Intelligence" (2005) 12 International Journal of Information Theories & Applications 277 at 277.

[40] Peter Vamplew et al, "Human-aligned artificial intelligence is a multiobjective problem" (2018) 20 Ethics and Information Technology 28 at 28.

[41] Treasury Board Secretariat, "Directive on Automated Decision-Making" (Ottawa: Her Majesty the Queen in Right of Canada, represented by the President of the Treasury Board, 2019) at Appendix [Directive on Automated Decision-Making]. Importantly, Canada does not at present define AI explicitly in any formal federal statutory law. In a 2020 recommendation, the Office of the Privacy Commissioner of Canada proposes amending the federal *Personal Information Protection and Electronic Documents Act* (PIPEDA) to address regulatory challenges raised by AI. See Office of the Privacy Commissioner of Canada, "A Regulatory Framework for AI: Recommendations for PIPEDA Reform" (Ottawa: Her Majesty the Queen in Right of Canada, represented by The Privacy Commissioner of Canada). The office's recommendation document does not explicitly define AI, but notes that the field has "immense promise" and is likely to, among other things, enable the detection and analysis of medical images, improve energy efficiency, and deliver highly individualized learning for students. It is not clear how Parliament would define AI, if at all, in any future statute amending PIPEDA.

accurate, consistent, and interpretable decisions made pursuant to Canadian law."[42] It explicitly depends on an analogy between human and computerized intelligence.

b. Human-Defined Process

But this analogy model is not our only option for defining AI. Cognitive and computer scientist Marvin Minsky proposes that we think about AI according to the ability of a machine to discover and mechanize "problem-solving processes."[43] In explaining what ought to count as a problem-solving process, Minsky proposes a non-exhaustive list that includes game-playing, theorem-proving, and pattern recognition.[44] This approach, while clearly inspired by human cognitive capacities, does not depend on an explicit comparison between computers and human beings. At least one influential textbook also takes something resembling this approach, describing AI as "the study of agents that receive percepts from the environment and perform actions."[45] On this view, the most important feature of AI appears not to be that it is concerned with computer systems that resemble or replicate human intelligence, but that it is concerned with computers capable of engaging in a certain kind of human-defined process. This is a kind of view reflected in two recent regulatory innovations in AI. In a 2019 recommendation for example, the Organisation for Economic Co-operation and Development (OECD) expresses the following:

> An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.[46]

---

[42] Directive on Automated Decision-Making, *ibid* at 4.

[43] Marvin Minsky, "Steps Toward Artificial Intelligence" 49:9 (1961) Proceedings of the Institute of Radio Engineers 8 at 8 ("Along with the development of general-purpose computers, the past few years have seen an increase in effort toward the discovery and mechanization of problem-solving processes. Quite a number of papers have appeared describing theories or actual computer programs concerned with game-playing, theorem-proving, pattern-recognition, and other domains which would seem to require some intelligence").

[44] Minsky, *ibid*.

[45] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed (Hoboken: Pearson Education, 2003) at vii.

[46] OECD, *Recommendation of the Council on Artificial Intelligence*, *supra* note 15 at s 1.

This conception is adopted nearly in its entirety by the European Commission in a recent draft regulation on AI. Released in April 2021, the proposal refers to AI as software that can "for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."[47] The draft regulation outlines three specific techniques that could be used for the development of an AI model: machine learning, logic and knowledge-based approaches, and statistical approaches such as Bayesian estimation.[48] In defining AI on these terms, the European Commission explicitly aims to build a definition that is "as technology neutral and future proof as possible" while also being "narrow, clear and precise."[49] It is uncertain from our present vantage whether the OECD and European Commission will succeed in their ostensible mission of carving out a conception of AI that is capable of resisting the inevitable advancement of the technology. These mirroring definitions follow Minsky's reasoning and do not explicitly depend on a computer's ability to perform tasks in a manner resembling human cognition. They do not, in other words, rely on an analogy between intelligence exhibited by humans and functions carried out by computers. Instead, all that appears to matter in this process-focussed approach is (1) that a machine produces outputs according to a set of human-defined objectives and (2) that these outputs influence a real or virtual environment.

Neither of the kinds of AI definition I have outlined are perfect. Common in both is the general sense that AI is a field concerned with the study and development of complex computer systems that capable of carrying out abstract functions and providing certain informational or

---

[47] 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*, OJ, COM(2021) 206 final, 2021/0106(COD) at art 3 [*Proposal for AI Regulation*].

[48] *Proposal for AI Regulation*, *ibid* at Annex I ("Artificial Intelligence Techniques and Approaches referred to in Article 3, point 1: (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; (c) Statistical approaches, Bayesian estimation, search and optimization methods").

[49] *Proposal for AI Regulation*, *ibid* at ss 3.1 & 5.2.1.

decisional outputs. Whether we need to say more than this is not a debate resolvable in this essay. For our present purposes, the best we might be able to say is that AI's contours are not plainly demarcated and that defining the scope of the field is an enterprise as old as the field itself. I will in this essay generally adopt the process-driven definition of AI defended by Minsky, the OECD, and the European Commission. In doing so, it is possible that I am overestimating the conceptual difficulty of asserting a cogent analogy between intelligence as expressed in humans and certain sophisticated machine functions. Even if the conceptual difficulty may be resolved, I think the analogy raises a set of theoretical and philosophical questions that unnecessarily complicate the project of defining AI's scope. In any event, it might be more productive to instead simply describe the kinds of paradigmatic methods and processes that constitute examples of AI applications on most generally agreed definitions. Below, I do this by summarizing machine learning (ML), perhaps the most significant AI innovation in recent decades. I also briefly outline ML training, the mechanism according to which ML models engage with the data that powers them.

2. *Unexplainable AI*

In the scope of this essay, the most singularly important species of AI application is machine learning. Subsumed within the larger AI field, hardly any of AI's permutations have been the subject of as much fanfare and promise as ML.[50] Many of the most impressive and potentially disruptive innovations in AI have been brought about by ML. Self-driving cars,[51] targeted online advertising,[52]

[50] See e.g. Adam J Russak et al, "Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype" (2020) 25:5 Journal of Cardiovascular Pharmacology & Therapeutics 379.

[51] See Jack Stilgoe, "Machine learning, social learning and the governance of self-driving cars" (2018) 48:1 Social Studies of Science 25.

[52] See Neil Shah et al, "Research Trends on the Usage of Machine Learning and Artificial Intelligence in Advertising" (2020) 19:5 Augmented Human Research 18.

and deepfaked media[53] all depend heavily on ML programming. ML is the "method of choice" for developing practical software in fields as diverse as "computer vision, speech recognition, natural language processing, [and] robot control." [54] Its continuing evolution is likely to affect nearly every field of human endeavour, including, as I outline in later parts of the essay, the practice of medicine.[55] Though ML models are some of AI's most powerful and promising applications, many of them have a feature likely to be legally and ethically quite confounding: they are unexplainable. In this part of the chapter, I describe machine learning, its subsidiary category deep learning, and the unexplainable character of many of the systems employing these methods. I briefly preface how unexplainable models might pose problems for the law.

a. Machine Learning

ML is concerned principally with one aspect of computerized intelligence: a machine's capacity to learn.[56] I will not at this stage comment extensively on the conceptual nature of learning as I did on the nature of intelligence above. To be sure, this is not because there are no conceptual questions worth asking—far from it, an analogy between human and computer learning is likely just as theoretically dubious as one between human and computer intelligence—but rather because the formulation of AI we identified above might make much of this discussion unnecessary. Insofar as we think of AI in terms of a model's capacity to make predictions, recommendations, or decisions

[53] See Marie-Helen Maras & Alex Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos" (2019) 23:3 International Journal of Evidence & Proof 255.

[54] MI Jordan & TM Mitchell, "Machine Learning: Trends, Perspectives, and Prospects" (2015) 349:6245 Science 255 at 255.

[55] See Mei Chen & Michel Decary, "Artificial Intelligence in Healthcare: An Essential Guide for Health Leaders" (2020) 33:1 Healthcare Management Forum 10 ("It is evident that AI has begun to affect almost every aspect of healthcare, from clinical decision support at points of care, patient self-management of chronic conditions at home, to drug research in the real world. The development and deployment of AI technology, however, is challenging and costly").

[56] See Igor Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective" (2001) 23:1 Artificial Intelligence in Medicine 89 ("Artificial intelligence is a part of computer science that tries to make computers more intelligent. One of the basic requirements for any intelligent behavior is learning. Most of the researchers today agree that there is no intelligence without learning. Therefore, machine learning is one of major branches of artificial intelligence and, indeed, it is one of the most rapidly developing subfields of AI research").

according to human-defined objectives, we might straightforwardly think about computer learning as something like the improvement of a system with the respect to the same set of human-defined objectives. What it is for a computer system to learn, in other words, might simply be that it gets better at making the kinds of predictions, recommendations, or decisions that it was designed to make. Essential in ML is that models learn and improve automatically over time.[57] Whereas more conventional AI operates within a carefully defined set of parameters, executing only those functions it has been explicitly programmed to execute, an ML model is capable of successfully performing tasks it has not been directly programmed to perform.[58] This tendency permits ML models to complete tasks at a significantly higher level of abstraction and complexity than conventional AI, constrained as it is by the ability of a human programmer to direct a model's conduct in code. It is far beyond the scope of the present chapter to provide a complete technical accounting of the way ML systems operate, but two related points are worth addressing.

First, ML systems essentially learn to improve over time by generalizing from data.[59] In this respect, the recent proliferation of ML is tied unavoidably to the recent corollary proliferation of data from nearly every field of human activity. This trend, popularly described as Big Data, refers to the creation and increasing availability of structured and unstructured information from numerous sources.[60] This Big Data is the fuel on which ML systems feed. Second, this process of learning to generalize from data, often called model training, is commonly, though not always, supervised by a

[57] Jordan & Mitchell, *supra* note 54 at 255 ("Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?").

[58] Ethem Alpaydin, *Introduction to Machine Learning* (Cambridge: MIT Press, 2014) at 3–4.

[59] Danilo Bzdok, Naomi Altman & Martin Krzywinski, "Statistics versus Machine Learning" (2018) 15:4 Nature Methods 233 at 233–234.

[60] Andrea De Mauro, Marco Greco & Michele Grimaldi, "A Formal Definition of Big Data Based on its Essential Features" (2016) 65:3 Library Review 122; Amir Gandomi & Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics" (2015) 35:2 International Journal of Information Management 137.

human programmer.[61] An ML model programmed through supervised training is provided both sample input and output data. Output data is either annotated by a human programmer or is framed according to some objective categorization metric such that the model is capable of correctly generalizing training data to novel information inputs.[62] Unsupervised model training functions, in a sense, as the mirror image of this. Models are provided input data and are left to identify natural pattern clusters on their own.[63] This distinction might be difficult to clearly make out. Michael Jordan and Tom Mitchell describe a model trained to identify credit-card fraud, and which may be helpful in discriminating supervised from unsupervised training:

> In learning to detect credit-card fraud, the task is to assign a label of 'fraud' or 'not fraud' to any given credit-card transaction. The performance metric to be improved might be the accuracy of this fraud classifier, and the training experience might consist of a collection of historical credit-card transactions, each labeled in retrospect as fraudulent or not. Alternatively, one might define a different performance metric that assigns a higher penalty when 'fraud' is labeled 'not fraud' than when 'not fraud' is incorrectly labeled 'fraud.'[64]

For a hypothetical model tasked with evaluating credit-card transactions for indications of fraud, in other words, a supervised learning approach would require providing the system input data in the form of real or imagined credit-card transactions and labelled output data in the form of 'fraud' or 'not fraud' designations associated with input credit-card transactions. An ML model trained in this manner would, drawing on its experience assessing labelled transactions, presumably be capable of labelling future arbitrary transaction inputs with a relatively high degree of accuracy.[65] An

---

[61] See DP O'Reagan, "Putting Machine Learning into Motion: Applications in Cardiovascular Imaging" (2020) 75 Clinical Radiology 33 at 34 ("ML takes many forms but most relevant to cardiovascular imaging is two broad categories described as supervised and unsupervised learning. Supervised learning involves obtaining prior knowledge that is used to train the model, which often consists of human image annotation or objective categorisation").

[62] O'Reagan, *ibid* at 34.

[63] O'Reagan, *ibid* at 34 ("In contrast, unsupervised learning often involves searching for natural clusters in the data that may identify similar groups"); see also Jordan & Mitchell, *supra* note 54 at 258.

[64] Jordan & Mitchell, *supra* note 54 at 255.

[65] See Mir Henglin et al, "Machine Learning Approaches in Cardiovascular Imaging" (2017) 10:10 Circulation: Cardiovascular Imaging 1 at 3 ("Labeled data result from associating unlabeled data with one or more meaningful descriptions. A label may be the definition of a measurement, the definition of a clinical trait, or the definition of a clinical outcome. For instance, a linear measure may be labeled as LVEF, a binary variable may be labeled as denoting the presence

unsupervised approach, in contrast, operates only on unlabelled input data. In the case of an unsupervised model for identifying credit-card fraud, training data would consist only of a set of unlabelled real or hypothetical credit-card transactions. A model of this kind would be, in a sense, left alone to identify naturally occurring data clusters such that fraudulent transactions could accurately be distinguished from not fraudulent transactions. Importantly, while a model operating in this way organizes relevant transaction data into natural clusters, these clusters would be unlabelled: an unsupervised model would not be capable of identifying which clusters are composed of fraudulent or not fraudulent transactions.[66] Cluster labeling for unsupervised models is a subordinate process that occurs following data categorization. This might occur through manual intervention or secondary automated process. In any case, *ex post* category labeling is an advantage for unsupervised models, for labeling already organized categories is typically quicker and cheaper than labeling unorganized output data as part of supervised learning.

Whether to employ supervised or unsupervised model learning methods depends on an array of factors, including training data complexity and the kinds of resources available for model design.[67] While much of modern model training is supervised, alternative methods, including combined training, semi-supervised training, and reinforcement training are increasingly common.[68] Several

---

or absence of LV hypertrophy, and another binary variable may be labeled as denoting the presence or absence of heart failure. Labels for data are often obtained by asking humans to carefully analyze or make judgments about unlabeled data (eg, asking a technical expert to trace the LV endocardium in multiple views to derive a biplane Simpson LVEF or asking an expert over-reader to adjudicate the presence or absence of rheumatic mitral valve disease). Thus, the process of labeling data often incurs substantial time and resource costs. The most successful machine learning algorithms, namely supervised learning algorithms, all require labeled data").

[66] See O'Reagan, *supra* note 61 at 34 ("In contrast, unsupervised learning often involves searching for natural clusters in the data that may identify similar groups. The same concept applies to both: a ML algorithm is used to learn how to optimally classify or cluster data by minimising a pre-specified "loss function." For instance, if we want to label (segment) anatomical structures on an image the loss function might be a term that evaluates the concordance between the segmented image and the ground-truth labels").

[67] Jordan & Mitchell, *supra* note 54 at 258–259.

[68] See Adriano Pinto et al, "Combining Unsupervised and Supervised Learning for Predicting the Final Stroke Lesion" (2021) 69 Medical Image Analysis 1.

commenters suggest that unsupervised learning is likely to become the dominant approach to AI learning in years to come.[69] Most learning in the natural world, after all, is unsupervised: human beings and other animals do not learn exclusively through exposure to labelled input-output pairs from which generalizations might be drawn.[70] As computerized models become more sophisticated, the potential for machines to learn and improve in an unsupervised manner will likely balloon. I will have more to say on the relationship between learning models in biology and in computing below. For now, what is important for our purposes is just that each of these ML training methods address the challenge of developing computational models capable of improving automatically over time through a process of generalizing in a structured way from input data. ML models, irrespective of their training background, fundamentally operate by recognizing patterns embedded in potentially massive datasets. This is a relevant point in later sections of the essay, in which I consider the advanced pattern-recognition capacities of ML models as compared with human decision-makers. Below, I describe a critical feature of the kind of ML pattern-recognition introduced here: that much of the time, it is impossible to know how or why a model categorized input data as it did.

b. Unexplainable Models

Some of the most uniquely powerful ML models employ a computational structure not described in the section above: deep learning. Deep learning (DL) models are "making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years."[71] DL models are, in very broad terms, constructed of artificial neural networks, computational

[69] Yann LeCun, Yoshua Bengio & Geoffrey Hinton, "Deep Learning" (2015) 521 Nature 436 at 442 ("We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress") [LeCun et al].

[70] LeCun et al, *ibid* at 442 ("Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object").

[71] LeCun et al, *ibid* at 436.

systems essentially modeled on the structure of biological neurons.[72] These structures consist of multiple layers of code, a stack of simple modules designed not by human engineers, but learned by the model itself.[73] Subsequent neural layers extract increasingly complex, abstract features from input data, producing models capable of performing "extremely intricate functions" while being simultaneously highly "sensitive to minute details."[74] Consider a theoretical model trained to distinguish Samoyeds from white wolves. An ideally trained model would be capable of both distinguishing exceptionally similar visual cues—the fluffy white coat of a Samoyed from the ruffled white coat of a wolf—while ignoring irrelevant though equally minute cues such as variations in background, pose, lighting, and surrounding objects.[75] These capacities have made DL remarkably promising in many contexts, from cardiovascular imaging[76] to understanding and translating natural language,[77] areas in which conventional ML approaches are often inadequate.

DL approaches generate models that are phenomenally and notoriously complex.[78] Even the most straightforward DL models are typically composed of at least three layers of computation, and

---

[72] Anthony M Zador, "A Critique Of Pure Learning and what Artificial Neural Networks can Learn from Animal Brains" (2019) 10 Nature Communications 1 at 2.

[73] LeCun et al, *supra* note 69 at 438 ("The conventional option is to hand design good feature extractors, which requires a consider- able amount of engineering skill and domain expertise. But this can all be avoided if good features can be learned automatically using a general-purpose learning procedure. This is the key advantage of deep learning. A deep-learning architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings").

[74] LeCun et al, *ibid* at 438.

[75] LeCun et al, *ibid* at 438 ("Since the 1960s we have known that linear classifiers can only carve their input space into very simple regions, namely half-spaces sepa- rated by a hyperplane19. But problems such as image and speech recognition require the input–output function to be insensitive to irrelevant variations of the input, such as variations in position, orientation or illumination of an object, or variations in the pitch or accent of speech, while being very sensitive to particular minute variations (for example, the difference between a white wolf and a breed of wolf-like white dog called a Samoyed)").

[76] Karen Andrea Lara Hernandez et al, "Deep Learning in Spatiotemporal Cardiac Imaging: A Review Of Methodologies and Clinical Usability" (2021) 130 Computers in Biology & Medicine 1.

[77] Ronan Collobert et al, "Natural Language Processing (Almost) from Scratch" (2011) 12 Journal of Machine Learning Research 2493.

[78] See LeCun et al, *supra* note 69 at 438–439; Andrew L Beam & Isaac S Kohane, "Big Data and Machine Learning in Health Care" (2018) 319:13 JAMA 1317.

some are composed of far more. Many of the most intricate tasks performable by DL require somewhere between five and twenty layers of computation.[79] It is difficult to abstractly capture the significance of this. Each layer of computed representation produces magnitudes greater learning capacity, while also producing correspondingly greater abstraction and model complexity.

This complexity produces a distinctively salient effect: that DL models are usually unexplainable. Often described as the 'black box problem,' the multilayered complexity of some of our best DL models means that even expert human reviewers are incapable of understanding how the relevant system operates.[80] There are two broad senses in which we might mean that a DL model is unexplainable, inscrutable, or subject to a black box problem. In one respect, a DL model might be unexplainable insofar as it relies on categorization rules that are "too complex for us to explicitly understand."[81] A sufficiently targeted model applying explicit rules might organize data in surprising or unintuitive ways, accounting for factors no human reviewer would think relevant.[82] A model trained to assess a patient's risk of stroke might, for example, find that nectarine consumption is strongly correlated with negative health outcomes. Most human clinicians would view this is as a counterintuitive and probably erroneous conclusion. But in making this determination, it is possible the DL model is weighing correlated factors or measuring information proxies in a manner that, though highly predictive, is confusing or unclear to human observers. This does not mean that the model is not applying explicit and theoretically reviewable rules, but simply that the specific rules

[79] LeCun, *ibid* at 437 ("With multiple non-linear layers, say a depth of 5 to 20, a system can implement extremely intricate func- tions of its inputs that are simultaneously sensitive to minute details — distinguishing Samoyeds from white wolves — and insensitive to large irrelevant variations such as the background, pose, lighting and surrounding objects").

[80] Davide Castelvecchi, "Can we Open the Black Box of AI?" (2016) 538 Nature 21.

[81] Price *supra* note 12 at 430.

[82] See Jenna Burrell, "How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms" (2016) Big Data & Society 1 at 5–7 ("The primary purpose of this first example is to give a quick, visual sense of how the machine 'thinks.' Figure 4(a) should appear unintuitive, random, and disorganized. However, handwriting recognition specifically is not a 'conscious' reasoning task in humans either. Humans recognize visual elements in an immediate and subconscious way (thus there is certainly a kind of opacity in the human process of character recognition as well)").

and the way they are applied resist straightforward interpretation. We might call systems of this kind, with explicit but complicated rules, *shallowly unexplainable* models.[83]

In a second respect, and more strictly relevant for our purposes here, a model might be unexplainable insofar as the learning techniques it employs are literally unknowable: "no one, not even those who programmed the machine-learning process, knows exactly what factors go into the ultimate decisions."[84] In this case, it is not just that the model and its categorization rules are difficult to understand, they are *impossible* to understand. Given a sufficiently complex DL model, input data might be subject to multiple abstract functions such that the system sets out a categorization scheme on the basis of an arbitrarily large number of variables.[85] In a sense, this problem has the same contours as the first, though on a potentially much larger scale. In the case of a DL model that cannot be explained, it is not just that variables are being considered in causally unintuitive or surprising ways, but that we cannot even know for certain which of the variables are being taken into account and in what proportion to each other. Consider a DL model trained in image recognition. Suppose further that it has been trained on a vast number of labelled images: objects as diverse and "as random as zebras, fire trucks, and seat belts."[86] A model of this kind might be capable of correctly categorizing human faces as distinguishable from other kinds of images or objects, despite having never been trained on labeled images of human beings. Assuming that the relevant model is sufficiently

[83] Yavar Bathaee would call these *weakly* unexplainable models or weak black boxes. See Bathaee, *supra* note 14 at 906 ("The decision-making process of a weak black box are also opaque to humans. However, unlike the strong black box, weak black boxes can be reverse engineered or probed to determine a loose ranking of the importance of the variables the AI takes into account. This in turn may allow a limited and imprecise ability to predict how the model will make its decisions… weak black boxes may not entirely cause intent and causation tests to cease to function, though they still pose serious challenges for both legal doctrines").

[84] Price, *supra* note 12 at 430.

[85] Cynthia Rudin & Joanna Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition" (2019) 1:2 Harvard Data Science Review 1 at 2.

[86] Paul Voosen, "How AI detectives Are Cracking Open the Black Box of Deep Learning" (2017) Science Newsletter, online: <https://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning>.

sophisticated, it is likely the case that no one, not even the model's programmer, will be able to explain precisely how the model learned to identify human faces. This learned capacity is facilitated by artificial neuronal connections far too numerous and too complicated for a human reviewer to audit, much less to explain. As the programmer of an image recognition model of the sort described above notes: "We build amazing models…But we don't quite understand them. And every year, this gap is going to get a bit larger."[87] We might call systems of this kind *deeply unexplainable* models.[88] It is unclear precisely what proportion of DL models are deeply unexplainable, though it is likely that many of the most powerful and effective DL models are likely to be unexplainable in this sense.

An entire industry has emerged in recent years to address potential problems generated by unexplainable AI.[89] Computing methods largely aimed at reverse-engineering interpretable accounts of what a model did in reaching particular decisions have in recent years received significant attention.[90] But relatively little technical progress has been made in these efforts.[91] And in any event, success on these kinds of technical innovations might not provide the kinds of explanations that we want. Recent scholarship emphasizes a distinction between descriptive explanations of the way a model behaves on the one hand and normative evaluations about whether a model's decisions are

[87] Voosen, *ibid*.

[88] Yavar Bathaee would call these *strongly* unexplainable models or strong black boxes. See Bathaee, *supra* note 14 at 906 ("Strong black boxes are AI with decision- making processes that are entirely opaque to humans. There is no way to determine (a) how the AI arrived at a decision or prediction, (b) what information is outcome determinative to the AI, or (c) to obtain a ranking of the variables processed by the AI in the order of their im- portance. Importantly, this form of black box cannot even be analyzed ex post by reverse engineering the AI's outputs").

[89] See e.g. Carlos Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence" (2021) 34 Philosophy & Technology 265.

[90] Hani Hagras, "Toward Human-Understandable, Explainable AI" (2021) 51:9 Computer 28 at 30 ("[Explainable AI] is a DARPA program that is expected to enable 'third-wave AI systems,' in which machines understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real world phenomena. According to a 2016 DARPA report, the [explainable AI] concept provides an explanation of individual decisions, enables understanding of overall strengths and weaknesses, and conveys an understanding of how the system will behave in the future and how to correct the system's mistakes").

[91] Castelvecchi, *supra* note 80 at 21 ("Twenty-five years later, deciphering the black box has become exponentially harder and more urgent. The technology itself has exploded in complexity and application").

sensible on the other.[92] While computing methods aimed at explanation might be capable of providing a descriptive explanation—though as I note below even this might be doubtful—they are incapable of addressing the second-order question of whether the model's internal operations are normatively justifiable.[93] To the degree the law's evaluation of unexplainable AI depends on normative rather than purely descriptive accounts of a model's operation, this may be a significant problem. It might be so significant, in fact, that certain scholars argue against using unexplainable medical AI altogether.[94]

But prefacing an argument I make in greater detail in the third chapter below, it might be that challenges created by unexplainable AI are not substantially distinct. Noting that the world itself is complicated, certain authors argue that it is probably unsurprising that our best automated representations of complex phenomena will resist attempts at external definition and interpretation.[95] It may be that the best way of dealing with unexplainable AI then, particularly in medicine, will be to resist significantly inflating the problem relative to its effects on well-being. Human being, after all, carry an unexplainable model around with us all the time: "You use your brain all the time; you trust your brain all the time; and you have no idea how your brain works."[96] And in any event, the

[92] Andrew D Selbst & Solon Barocas, "The Intuitive Appeal of Explainable Machines" (2018) 87 Fordham Law Review 1085 at 1117 ("Notably, neither the techniques nor the laws go beyond describing the operation of the model. Though they may help to explain why a decision was reached or how decisions are made, they cannot address why decisions happen to be made that way. As a result, standard approaches to explanation might not help determine whether the particular way of making decisions is normatively justified").

[93] Marzyeh Ghassemi, Luke Oakden-Rayner & Andrew L Beam, "The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care" (2021) 3 Lancet Digital Health 745 at 748 ("Although most discussions and policies call for normative evaluations, current techniques are only capable of descriptive accounts and it is our own intuition that often 'serves as the unacknowledged bridge' between the two").

[94] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019) 1 Nature Machine Intelligence 206 at 206 ("Rather than trying to create models that are inherently interpretable, there has been a recent explosion of work on 'explainable ML,' where a second (post hoc) model is created to explain the first black box model. This is problematic. Explanations are often not reliable, and can be misleading, as we discuss below. If we instead use models that are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes").

[95] Castelvecchi, *supra* note 80 at 23 ("Ultimately, these researchers argue, the complex answers given by machine learning have to be part of science's toolkit because the real world is complex: for phenomena such as the weather or the stock market, a reductionist, synthetic description might not even exist").

[96] Castelvecchi, *ibid* at 23.

unexplainable character of our best AI models is not a bug. It is a feature of highly sophisticated automated decision-making. That a model is capable of drawing functional inferences between data points that human observers are incapable of identifying, much less understanding, is one of DL's "principal advantages" over more traditional AI.[97] Models capable of drawing out unintuitive, subtle inferences will consider factors and variable permutations most humans would ignore. An unexplainable DL model, in other words, might work to fill gaps in human reasoning and, in consequence, enhance rather than diminish our rational capacities. In any event, it is certain that the unexplainable character of certain ML systems will produce certain challenges. In the next chapter of the thesis, I describe what some of those challenges might be in the healthcare context. But first, I give an overview of some of the ways DL models are tangibly being used in medicine.

3. *Deep Learning in Canadian Medicine*

Nowhere is unexplainable AI likely to be more profoundly impactful in the medium-term than in medicine.[98] In this part of the chapter, I briefly outline some of the ways in which DL models are likely to be adopted in Canadian healthcare. While general AI methods have sweeping potential applications in hospital management, workflow optimization, and population health monitoring,[99] many of the most significant patient-facing applications of AI in medicine rely on DL methods.[100] DL

---

[97] Zednik, *supra* note 89 at 267–268 ("Thus, in contrast to their colleagues working in other approaches in AI, ML developers do not actually specify how an AI problem is solved but merely specify the conditions under which a solution may eventually be found. This relative lack of influence on the way in which problems are actually solved is one of the principal advantages of Machine Learning over traditional approaches in AI").

[98] See generally, Michael da Silva, *AI in Health Care: A Fusion of Law & Science* (CIFAR: Toronto, 2021); Fei Wang, Lawrence Peter Casalino & Dhruv Khullar, "Deep Learning in Medicine: Promise, Progress, and Challenges" (2018) 179:3 Health Care Reform 293.

[99] Chen & Decary, *supra* note 55 at 14.

[100] Travers Ching et al, "Opportunities and obstacles for deep learning in biology and medicine" (2018) 15 Journal of the Royal Society Interface 1 at 2.

models are, for example, intimately engaged in diagnosis and patient care, potentially exposing them to unique challenges and producing direct effects on individual health outcomes. Medical imaging is likely the most prolific present use case for DL models in the practice of medicine, with certain models performing at least as well, and often better, than human clinicians in the identification of disease from scan images.[101] These effects have been especially pronounced in oncology and cardiovascular imaging, in which advanced DL models promise to augment the effectiveness of human physicians in the interpretation of nuclear imaging, cardiac magnetic resonance imaging (MRI), and computed tomography (CT).[102] Owing to significant academic optimism in the promise of DL to revolutionize medicine, it is prudent to understand and document the degree to which DL models are presently incorporated into medical practice. Following the excellent work of Stan Benjamens and colleagues documenting the approval of AI models by the United States Food and Drug Administration (FDA),[103] I sought to understand whether potentially unexplainable DL models have been approved for clinical use in Canada.

On reviewing Health Canada medical device approvals, I found at least ten DL models have been approved for clinical use in Canada, with most of them performing medical imaging functions. I identified these DL medical models by performing an iterative search of Health Canada's Medical Devices Active Licence Listing, a reference tool documenting basic information about medical devices presently approved by Health Canada's Medical Devices Bureau.[104] I performed company name and device name searches using fixed search terms 'AI,' 'artificial intelligence,' 'deep

[101] Wang et al, *supra* note 98 at 293.

[102] KR Siegersma et al, "Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist" (2019) 27 Netherlands Heart Journal 403; Andrew Daniel Trister, "The Tipping Point for Deep Learning in Oncology" (2019) 5:10 JAMA Oncology 1429.

[103] See Stan Benjamens, Pranavsingh Dhunnoo & Bertalan Meskó, "The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database" (2020) 3 Nature Digital Medicine 1 at 3–4.

[104] Active Licences Search, *supra* note 18 ("MDALL contains product-specific information on all medical devices that are currently licensed for sale in Canada, or have been licensed in the past").

learning,' 'deep,' 'smart,' and 'software.' I additionally cross-referenced search results with manufacturer public statements, white papers, and news releases to determine whether approved models employ DL methods. I excluded approved models for which the device manufacturer does not claim explicitly to have employed DL methods in the approved model's initial programming. I further cross-referenced these results with an online database of FDA-approved AI models maintained by Benjamens and colleagues to better assess model function and technical orientation. This approach has several methodological limitations. First, results outlined in the figure below should not be interpreted to be an exhaustive list of approved medical devices employing DL modelling and methods in Canada. Health Canada does not maintain a publicly accessible accounting of software categorization for approved medical devices. It is possible that other DL models have been approved and are not captured in the results below because (1) the relevant company or device name is not related directly to AI or DL techniques, (2) the model is approved by Health Canada but has not received FDA approval, (3) the model does not require Health Canada approval because it falls into one or more of the regulator's explicit regulatory exclusion criteria,[105] or (4) the relevant manufacturer does not publicly state their use of DL methods in model design. Second, and as a corollary of this latter point, model manufacturers may publicly misrepresent their use of DL methods, resulting in the inclusion of models that do not, strictly speaking, actually employ DL. In the table below, I have identified instances in which a manufacturer publicly states multiple or more specific AI methods. Third, that a particular model has been approved by Health Canada does not necessarily definitively demonstrate that the technologies are in active clinical use. But the availability of these models for clinical application nevertheless provides a snapshot of the DL medical model landscape in Canada and demonstrates that the tools outlined below are at least susceptible to use in patient care.

[105] See e.g., Health Canada, "Guidance Document: Software as a Medical Device (SaMD): Definition and Classification" (Ottawa: Her Majesty the Queen in Right of Canada, as Represented by the Minister of Health, 2019) at 9–10.

Table: Example Deep Learning Models Approved for use in Canada

Device name	Manufacturer	AI approach	Description	Health Canada device identifier	First issue date
Icobrain	Icometrix NV: Leuven, Belgium	Deep Learning	"Quantify disease-specific brain structures for acute and chronic neurological conditions on MR and CT."[106]	B621ICO BRAIN30	2018-04-26
Guardian™ Connect App	Medtronic Minimed: Northridge, California	General AI, Machine Learning	"Provides continuous, real-time trend information about glucose levels for people with diabetes. It allows for appropriate intervention (after verifying with a blood fingerstick test) to mitigate hyperglycemia (high blood sugar) or hypoglycemia (low blood sugar), maximizing the patient's time in the optimal glucose target range."[107]	CSS7200	2018-05-09
Arterys Mica Software	Arterys, Inc: San Francisco, California	General AI, Deep Learning	Deep Learning with cloud computing for automatic quantification and segmentation of ventricles from CT and MRI; additional applications in cancer, liver, and lung disease.[108]	AMM6	2018-10-17
Deep Learning Image Reconstruction	GE Healthcare Japan Corporation: Tokyo, Japan	Deep Learning	"GE Healthcare's deep learning image reconstruction (DLIR) is the first Food and Drug Administration (FDA) cleared technology to utilize a deep neural network-based recon engine to generate high quality TrueFidelity	5835944	2020-07-17

[106] Icometrix, "Enabling value-based care for people with neurological conditions" (2021), online: <https://icometrix.com>.

[107] Medtronic, "Guardian™ Connect Continuous Glucose Monitoring (Cgm) System Now Licensed in Canada for People Living with Diabetes" (May 23, 2018), online: <https://www.medtronic.com/ca-en/about/news/Guardian_Connect_Press_Release.html>.

[108] Arterys, "About Us" (2021), online: <https://arterys.com/about-us>.

			computed tomography (CT) images."[109]		
Critical Care Suite	GE Medical Systems, LLC: Waukesha, Wisconsin	General AI, Deep Learning	"Quickly identify and help prioritize critical cases such as Pneumothorax" from x-ray images.[110]	5830234	2020-07-22
ProFoundAI™	iCAD, Inc: Nashua, New Hampshire	Deep Learning	"ProFound AI™ [uses] deep learning technology [and] is intended to be used concurrently by radiologists while reading digital breast tomosynthesis (DBT) exams."[111]	D70177	2020-08-05
AI-RAD Companion (Muscoskeletal)	Siemens Healthcare GMBH: Erlangan, Germany	Deep Learning	Muscoskeletal CT image reconstruction.[112]	11270067	2020-12-24
AI-RAD Companion (Cardiovascular)	Siemens Healthcare GMBH: Erlangan, Germany	Deep Learning	Cardiovascular CT image reconstruction.[113]	11270066	2021-01-14
AI-RAD Companion (Pulmonary)	Siemens Healthcare GMBH: Erlangan, Germany	Deep Learning	Pulmonary CT image reconstruction.[114]	11270065	2021-01-14
Advanced Intelligent Clear-IQ Engine (AiCE) for MR	Canon Medical Systems Corporation: Otawara-Shi, Japan	Deep Learning, Deep Convolutional Neural Network	"MR [magnetic resonance] Deep Learning reconstruction technology."[115]	MSSW-DLR1	2021-06-01

[109] Jiang Hsieh et al, "A new era of image reconstruction: TrueFidelity™ (Technical White Paper, JB68676XX, 2019).

[110] GE Healthcare, "Critical Care Suite 2.0" (2021), online: <https://www.gehealthcare.com/products/radiography/mobile-xray-systems/critical-care-suite-on-optima-xr240amx>.

[111] iCAD, "Artificial Intelligence for Digital Breast Tomosynthesis: Reader Study Results" (White Paper, DMM253 Rev B, 2020).

[112] Siemens Healthineers, "AI-Rad Companion" (2021), online: <https://www.siemens-healthineers.com/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion>.

[113] Siemens Healthineers, *ibid*.

[114] Siemens Healthineers, *ibid*.

[115] Canon Medical Systems, "Advanced intelligent Clear-IQ Engine (AiCE)" (2020), online: <https://global.medical.canon/products/magnetic-resonance/aice>.

Several conclusions might be drawn from the results above. For one thing, and most significantly, it is exceptionally likely that DL models are being presently applied in the healthcare context in Canada. Applications of DL modelling in medicine is not some theoretical objective, a concern for future generations of clinicians and commenters: these systems are already here, affecting physician practice and patient care. For another thing, most of the DL presently approved in Canada appears to function in medical imaging. With the exception of only one of the models outlined above, the Guardian™ Connect App, each of the reviewed DL models assists in the clinical imaging of various tissues. This is perhaps not unsurprising considering DL's uniquely proficient capacities in image recognition. It is nevertheless worth underlining, for it helps us to better understand exactly what DL models are doing in the context of Canadian healthcare. To be sure, that the models outlined above adopt DL methods does not, on its own, directly imply that they are unexplainable, which is the primary concern of this part of the chapter. Rather, it simply indicates that they *probably* are. Health Canada does not record whether approved AI models are unexplainable in the sense I outlined above, but recall that our best DL models, those trained to perform complex tasks such as image recognition, typically cannot be explained, even by their programmers. This is a question to which I will return in subsequent portions of the thesis. For now, it may be sufficient to say that the ten DL models presently approved by Health Canada are probably technically unexplainable, at least in respect of some of their functions.

Conclusion

In this chapter, I set out to do three things. First, I sought to provide a general overview of the field of AI. Second, I sought to summarize the manner in which certain applications of AI are likely to be technically unexplainable. Third, I sought to provide some concrete examples of the use of potentially unexplainable AI in Canada's healthcare system. Each of these themes will help to motivate this essay's second chapter, in which I consider the implications in law of the adoption of unexplainable AI in healthcare.

CHAPTER TWO: IMPLICATIONS for the PRACTICE and REGULATION of MEDICINE

This second chapter surveys some of the potential implications of unexplainable AI for the practice and regulation of medicine. It suggests that the unexplainable character of many of our best medical AI is in tension with the traditional operation of malpractice law. Unexplainable medical models may interfere with two foundational elements of the law of clinician civil liability: fault and causation. This interference may in turn generate two distinct challenges for medicine: (1) that clinicians are unlikely to have a clear sense of their obligations with respect to the use of unexplainable models in patient care and (2) that injured patients may find that their capacity to seek redress in the law of civil obligations is markedly restricted by the unexplainable character of medical interventions used in their care. While neither of these factors, in my view, constitute an overriding objection to the clinical use of unexplainable AI, they help to contextualize how these technologies might pose a conceptual challenge to our usual ways of thinking about malpractice law. More to the point, these challenges help to foreground discussion surrounding rights to explanation, which is the subject of the third chapter below. I suggest in this chapter that while unexplainable AI generates genuine difficulty for medicine, these challenges are not in principle insurmountable.

Malpractice law plays a considerable role structuring the practice of medicine.[116] It influences how clinicians behave[117] and functions as perhaps the most important frame for understanding the legal significance of the clinician-patient relationship.[118] And it does this even while being ineffective

[116] Lorian Hardcastle, "Medical Negligence Law in Canada" in Joanna Erdman. Vanessa Gruben & Erin Nelson, eds, *Canadian Health Law & Policy*, 5th Edition (Toronto: LexisNexis, 2017) 305 at 305–3-6; Elaine Gibson, "Is It Time to Adopt a No-Fault Scheme to Compensate Injured Patients?" (2016) 47:2 OLR 303 at 309; Ronan Avraham & Max M Schazenbach, "Medical Malpractice" in Francesco Parisi, ed, *Oxford Handbook of Law & Economics, Volume II: Private & Commercial Law* (Oxford: Oxford University Press, 2017) 120–147 at 122.

[117] See e.g. Lee Black, "Effects of Malpractice Law on the Practice of Medicine" (2007) 9:6 American Medical Association Journal of Ethics 437; Erik Renkema, Manda Broekhuis & Kees Ahaus, "Conditions that influence the impact of malpractice litigation risk on physicians' behavior regarding patient safety" (2014) 14:38 BMC Health Services Research 1; Barry S Schifrin & Wayne R Cohen, "The Effect of Malpractice Claims on the Use of Caesarean Section" (2013) 27 Best Practice & Research Clinical Obstetrics & Gynaecology 269.

[118] See generally, Shawn HE Harmon, David E Faour & Noni E MacDonald, "Physician Dismissal of Vaccine Refusers: A Legal and Ethical Analysis" (2020) 13:2 McGill JL & Health 255; Claudia E Haupt, "Governing AI's Professional Advice"

at accomplishing one of its ostensibly primary goals: reducing patient injury.[119] While clinician-patient relationships are in both of Canada's dominant legal tradition derived from the law of contracts, redress for patient injury in the common law is addressed primarily through the law of torts.[120] This chapter uses medical malpractice in Canada's dominant private law traditions to understand some of the conceptual challenges potentially posed unexplainable medical AI. For these purposes, I refer variously to medical malpractice, clinician civil liability, and liability for clinical fault. I generally apply these terms and their permutations interchangeably.

While Canada's common law and Quebec's civil law apply distinct legal frameworks, each using their own language and making independent assumptions, they essentially structure liability in the medical context according to broadly similar constitutive factors. Canada's common law jurisdictions, for example, approach medical malpractice as an application of the law of torts, permitting an injured plaintiff can succeed on proving four essential elements: a duty of care, breach, injury, and causation. A plaintiff advancing a tort law claim for medical malpractice must prove that a defendant's breach of a duty of care caused a legally cognizable injury, that "but for the tortious conduct of the defendant, the plaintiff would not have sustained the injury complained of."[121]

(2019) 64:4 MLJ 665 at 677–678; Lorian Hardcastle & Colleen M Flood, "The Future of Health Law: A View Forward from 2016" (2016) Ottawa LR 299.

[119] Colleen Flood & Brian Thomas, "Canadian Medical Malpractice Law in 2011: Missing the Mark on Patient Safety" (2011) 86:3 Chicago–Kent L Rev 1053 at 1054 ("We argue that tort law, as it stands in 2011, misses the mark in addressing the hidden epidemic in patient safety; although we admit the paucity of robust empirical evidence makes it difficult to know whether rates of iatrogenic injury are worsening, stable or improving. There is, however, rising concern regarding the quality of care and safety of patients in privately financed and informal health care settings, e.g. in private clinics, in long-term care homes and in home care. This suggests that what we do know about the rates of adverse events in the hospital setting may be merely the tip of the iceberg"); Habiba Nosheen & Andrew Culbert, "As Fewer Patients Sue their Doctor, the Rate of Winning Malpractice Suits is Dropping too" (2019) CBC News, online: <https://www.cbc.ca/news/health/medical-malpractice-doctors-lawsuits-canada-1.4913960> ("An analysis of the past 40 years shows that as the number of doctors increased, the rate of patients suing has dropped. For the cases that do make their way to court, the number of patients who have won has also gone down").

[120] Hardcastle, *supra* note 116 at 307; Ernest J Weinrib, *Tort Law: Cases & Materials*, 5th Edition (Toronto: Emond Montgomery Publications, 2019) at 51–54; Suzanne Philips-Nootens, Robert P Kouri & Pauline Lesage-Jarjoura, *Éléments de reponsabilité civile médicale: le droit dans le quotidien de la médecine*, 4th edition (Montreal: Éditions Yvons Blais, 2016) at 48.

[121] *Snell v Farrell*, [1990] 2 SCR 311 at 320, 72 DLR (4th) 289.

Quebec's civil law addresses medical malpractice primarily in view of the contractual nature of the relationship between clinicians and their patients.[122] Article 1458 of the *Civil Code* notably outlines the foundational elements of liability of the law of contracts in Quebec: anyone who fails to abide by their contractual undertakings, and in so doing injures another, is obliged to remedy the injury.[123] Though it is an essential element of malpractice law in both common and civil law traditions, I do not focus directly on the element of injury in these kinds of claims. I assume for the purposes of the present chapter that a plaintiff's capacity to prove injury is not likely to be affected significantly by the adoption of unexplainable AI in medicine. To be sure, unexplainable AI might have serious implications for injury in other civil liability contexts: privacy harms that might become more widespread with the use of unexplainable models might, for example, be uniquely difficult to prove in litigation.[124] But in medical malpractice, much of the conceivable injury to which patients are likely to be subject will not reasonably be contestable. Any court would, for example, find that illness progression following misdiagnosis is an injury, even if the circumstances in which it occurs do not admit of recovery in civil liability. But the remaining elements of conventional medical malpractice claims—failure to meet an obligation imposed by law and the causation of injury—are sure to be

[122] See e.g. Philips-Nootens et al, *supra* note 120 at 2 ("Les relations juridiques entre le médecin et son patient peuvent naître d'un contrat ou relever, en l'absence de contrat, d'obligations imposées par le législateur. Si la plupart des obligations se retrouvent dans ces deux types de situations, chacune d'elles comporte néanmoins des particularités quant à un éventuel recours en responsabilité").

[123] See art 1458 CCQ ("Every person has a duty to honour his contractual undertakings. Where he fails in this duty, he is liable for any bodily, moral or material injury he causes to the other contracting party and is bound to make reparation for the injury; neither he nor the other party may in such a case avoid the rules governing contractual liability by opting for rules that would be more favourable to them").

[124] See M Ryan Calo, "The Boundaries of Privacy Harm" (2011) 86 Indiana LJ 1131 at 1132 ("A burn is an injury caused by heat. It has symptoms. It admits of degrees. When a doctor diagnoses a burn, she immediately gains insights into how best to treat it. She can rule out other causes. She can even make recommendations on how to avoid this particular harm in the future. What is a privacy harm? What makes it distinct from a burn or some other harm? We are often at a loss to say. Privacy harm is conceptualized, if at all, as the negative consequence of a privacy violation. Far from a source of leverage or insight, privacy harm often operates as a hurdle to reform or redress"); Daniel J Solove & Danielle Keats Citron, "Privacy Harms" (2021) GWU Legal Studies Research Paper No. 2021-11 at 3 ("Law's treatment of privacy harms is a jumbled, incoherent mess. Countless privacy violations are left unaddressed because courts refuse to recognize harm that has been suffered").

more contentious. This chapter outlines how this is so. First, I briefly consider the question of obtaining consent for the use of unexplainable medical AI. I suggest that the fundamental clinical obligation to obtain the informed consent of individuals subject to a clinical intervention might be complicated by the application of medical devices no one can robustly explain. Second, I argue that unexplainable medical AI poses a challenge for jurists understanding whether a clinician's conduct in using an unexplainable device constitutes a fault for the purposes of medical malpractice law. I suggest that unexplainable medical AI makes applying the dominant fault standard in both common and civil law traditions more conceptually complicated. Third, I consider causation as an element of malpractice liability and suggest that unexplainable AI will probably make it more difficult for jurists to understand how, as a matter of fact and law, patient injuries in the clinical adoption of unexplainable AI were caused. I argue that these difficulties are likely to have a considerable and uncertain impact on the practice and regulation of medicine.

1. *Obtaining Consent*

Imagine that a clinician, Dr. *x*, uses a sophisticated DL model, Cardio *y* in their cardiology practice. Cardio *y* reads MR cardiovascular images and assists physicians in the diagnosis of ventricular disorder. Cardio *y* was recently approved as a medical device by Health Canada and the best available evidence suggests that the system is at least as effective as human physicians in most cardiovascular diagnosis applications. But no one, not even Cardio *y*'s manufacturer, knows for sure how it works. Dr. *x* successfully uses the model to categorize dozens of patients and it quickly becomes an indispensable tool in their practice. Dozens of cardiology specialists across the country start using Cardio *y*, with only a small number of documented instances of misdiagnosis. Several publications appear in leading specialty and generalist medical journals that give additional support for Cardio *y*'s safety and efficacy. Imagine that a patient, Patient *z*, presents to Dr. *x* with symptoms suggesting a severe ventricular abnormality. Dr. *x* wishes to assess Patient *z*'s condition by using Cardio *y* and sets

out to obtain their consent to do so. Patient *z* has questions about Cardio *y*: what it does, how it works, how it differs from other diagnostic instruments. Dr. *x* does their best to address Patient *z*'s concerns, but on technical questions about how the model functions, can gesture only vaguely at an answer. Dr. *x* says something to the effect of "Cardio *y* is an approved medical device. It has been determined by Health Canada to be safe and effective." All of this is true, if unilluminating.

Patient *z* agrees, somewhat reluctantly, to follow Dr. *x*'s advice. Informed by a classification report generated by Cardio *y*'s MR image processing, Patient *z* is diagnosed with a rare and complex ventricular abnormality. Dr. *x* orders a course of treatment without first doing additional tests or explicitly considering alternate diagnoses. Patient *z*'s condition soon worsens. After a rapid decline, Patient *z* dies. As it happens, their condition had been misdiagnosed. Patient *z*'s estate sues Dr. *x* for malpractice, alleging in part that Dr. *x* failed to obtain Patient *z*'s informed consent prior to using Cardio *y* in their care. In not fully explaining how Cardio *y* works, Patient *z*'s estate alleges that the relevant treatment ought to be remedied in medical malpractice. In this first section, I briefly outline the duty of clinicians to obtain informed consent. I observe that the unexplainable character of newly approved medical devices creates a potential problem for obtaining informed consent.

a. Obligation to obtain informed consent

The obligation of clinicians to obtain the informed consent of their patients for medical intervention is perhaps their most firmly established and foundational professional duty. Our modern conception of the duty to obtain informed consent has roots in the Nuremberg Trials, particularly in the conviction of twenty-three Nazi physicians for their involvement in horrifying human experimentation and mass murder.[125] Nuremberg helped to establish a duty on the part of medical professionals to obtain

[125] See Jochen Vollmann & Rolf Winau, "Informed consent in human experimentation before the Nuremberg code" (1996) 313 British Medical Journal 1445 at 1445 ("The issue of ethics with respect to medical experimentation in Germany during the 1930s and 1940s was crucial at the Nuremberg trials and related trials of doctors and public health officials. Those involved in horrible crimes attempted to excuse themselves by arguing that there were no explicit rules governing medical research on human beings in Germany during the period and that research practices in Germany were not different from those in allied countries. In this context the Nuremberg code of 1947 is generally regarded as the first document to set out ethical regulations in human experimentation based on informed consent"); See also *United States v Karl Brandt et al*

voluntary, competent, and informed consent for research. Cases decided in the United States and Canada over the course of the Twentieth Century transposed the Nuremberg duty into clinical care.[126] Clinicians in Canada are bound by a patchwork of legislative, jurisprudential, and ethical obligations to obtain the informed consent of their patients before carrying out an intervention. Quebec's *Civil Code*, for example, prohibits medical care conducted absent patient consent.[127] Anyone older than fourteen is entitled to consent on their own behalf while incapable adults and minors younger than fourteen are subject to specific alternative requirements, such as proxy consent provided by a parent or tutor.[128] These rules are supplemented in regulation, such as in the *Code of Ethics of Physicians*, which provides that patient consent must be "free and enlightened."[129] It further specifies that clinicians are bound to provide information about the nature, purpose, and possible consequences of a treatment intervention.[130] Ontario sets out broadly similar rules in the *Health Care Consent Act*, indicating that patient consent must be informed, voluntary, and obtained without misrepresentation

(Doctor's Trial, 1947), in *Records of the United States: Nuremberg War Crimes Trial* (Washington: National Archives and Record Service, 1974).

[126] See e.g. *Salgo v Leland Stanford*, (1957) 154 Cal App 2d 560, 317 P2d 170 ("A physician violates his duty to his patient and subjects himself to liability if he withholds any facts which are necessary to form the basis of an intelligent consent by the patient to the proposed treatment. Likewise the physician may not minimize the known dangers of a procedure or operation in order to induce his patient's consent. At the same time, the physician must place the welfare of his patient above all else and this very fact places him in a position in which he sometimes must choose between two alternative courses of action"); *Halushka v University of Saskatchewan*, [1965] 53 DLR 2nd 436; *Hopp v Lepp*, [1980] 2 SCR 192, 13 CCLT 66 ("In summary, the decided cases appear to indi-cate that, in obtaining the consent of a patient for the performance upon him of a surgical operation, a surgeon, generally, should answer any specific questions posed by the patient as to the risks involved and should, without being questioned, disclose to him the nature of the proposed opera-tion, its gravity, any material risks and any special or unusual risks attendant upon the performance of the operation").

[127] Art 11 CCQ ("No person may be made to undergo care of any nature, whether for examination, specimen taking, removal of tissue, treatment or any other act, except with his consent").

[128] See art 14 CCQ ("A minor 14 years of age or over, however, may give his consent alone to such care"); see also art 15 CCQ ("Where it is ascertained that a person of full age is incapable of giving consent to care required by his or her state of health and in the absence of advance medical directives, consent is given by his or her mandatary, tutor or curator").

[129] *Code of Ethics of Physicians*, CQLR, c M–9, r 17, s 28 ("A physician must, except in an emergency, obtain free and enlightened consent from the patient or his legal representative before undertaking an examination, investigation, treatment or research").

[130] *Code of Ethics of Physicians*, *ibid*, s 29 ("A physician must ensure that the patient or his legal representative receives explanations pertinent to his understanding of the nature, purpose and possible consequences of the examination, investigation, treatment or research which he plans to carry out").

or fraud.[131] The *Act* notes that consent to care is informed only if the patient receives the kind of information a reasonable person would require to make a decision in similar circumstances and the patient's requests for additional information are addressed.[132]

Professional bodies across the country contextualize and refine provincial statutory obligations to obtain informed consent in an array of policy documents and ethics statements. The Canadian Medical Association, for example, notes in its *Code of Ethics and Professionalism* that medical decision-making guided by the informed consent process is an ideally deliberative process "informed by the patient's experience and values and the physician's clinical judgment."[133] Clinicians are counseled to empower their patients to make informed decisions, among other things by communicating "material risks and benefits" and advising on the available "reasonable therapeutic options."[134] Canadian courts have since at least the 1980s also weighed in extensively on the rules surrounding clinical consent. In the landmark *Reibl v. Hughes* case, for example, the Supreme Court sets out a common law rule that clinicians are bound to disclose the risks a "reasonable person in the patient's position" would require to decide whether to accept or decline treatment.[135] This is a position

[131] *Health Care Consent Act*, RSO 1996, c 2, Sched A, s 11(1) ("The following are the elements required for consent to treatment: (1) the consent must relate to the treatment, (2) the consent must be informed, (3) the consent must be given voluntarily, (4) the consent must not be obtained through misrepresentation or fraud").

[132] *Health Care Consent Act*, *ibid*, s 11(2) ("A consent to treatment is informed if, before giving it, (a) the person received the information about the matters set out in subsection (3) that a reasonable person in the same circumstances would require in order to make a decision about the treatment; and (b) the person received responses to his or her requests for additional information about those matters").

[133] Canadian Medical Association, *Code of Ethics and Professionalism*, Ottawa: CMA, 2018, p 4 ("Medical decision-making is ideally a deliberative process that engages the patient in shared decision making and is informed by the patient's experience and values and the physician's clinical judgment. This deliberation involves discussion with the patient and, with consent, others central to the patient's care (families, caregivers, other health professionals) to support patient-centred care").

[134] Canadian Medical Association, *ibid*, p 5 ("Empower the patient to make informed decisions regarding their health by communicating with and helping the patient (or, where appropriate, their substitute decision-maker) navigate reasonable therapeutic options to determine the best course of action consistent with their goals of care; communicate with and help the patient assess material risks and benefits before consenting to any treatment or intervention").

[135] *Reibl v Hughes*, [1980] 2 SCR 880, 114 DLR (3d) 1 at 899 ("So too, other aspects of the objective standard would have to be geared to what the average prudent person, the reasonable person in the patient's particular position, would agree to or not agree to, if all material and special risks of going ahead with the surgery or foregoing it were made known to him. Far from making the patient's own testimony irrelevant, it is essential to his case that he put his own position forward").

further elaborated in subsequent decisions, such as in *Starson v. Swaze*, where the Court stipulates that informed consent, especially the capacity to refuse unwanted medical treatment, is "fundamental to a person's dignity and autonomy."[136] In the 2013 case *Cuthbertson v. Rasouli*, the Court found that the autonomy interests of patients "has historically been viewed as trumping all other interests, including what physicians may think is in the patient's best interests."[137] The *Rasouli* decision reaffirms the common law rule expounded in *Hughes*, that clinicians have an obligation to obtain consent that is voluntary and informed: clinicians must disclose the nature of a proposed treatment, including its risks, benefits, and alternatives.[138] This, in broad strokes is what the law of informed consent requires in Canada. Patient autonomy requires that no one be subject to medical intervention without first understanding precisely what it is they signing up for. Clinicians have an obligation to facilitate this process. It is a duty clearly established in statute, elucidated in professional ethics guidance, and affirmed in case law. And it is a duty likely made more complicated by unexplainable medical AI. I describe what the challenge might be in the section below.

b. Consenting to an unexplainable intervention

To the degree informed consent law requires clinicians to disclose the nature of medical interventions to which patients are subject, unexplainable medical AI encounters a problem. That a model's internal workings are unknowable suggests that patients may be unable to adequately assess the risks and

---

[136] *Starson v Swaze*, 2003 SCC 32, [2003] 1 SCR 722 at para 75 ("The right to refuse unwanted medical treatment is fundamental to a person's dignity and autonomy. This right is equally important in the context of treatment for mental illness: see Fleming v. Reid (1991), 4 O.R. (3d) 74 (C.A.), per Robins J.A., at p. 88: 'Few medical procedures can be more intrusive than the forcible injection of powerful mind-altering drugs which are often accompanied by severe and sometimes irreversible adverse side effects'").

[137] *Cuthbertson v Rasouli*, 2013 SCC 53, [2013] 3 SCR 341 at para 19 ("The patient's autonomy interest — the right to decide what happens to one's body and one's life — has historically been viewed as trumping all other interests, including what physicians may think is in the patient's best interests").

[138] *Cuthbertson v Rasouli*, *ibid* at para 18 ("The physician cannot override the patient's wishes to be free from treatment, even if he believes that treatment is in the vital interests of the patient. The patient's consent must be given voluntarily and must be informed, which requires physicians to ensure the patient understands the nature of the procedure, its risks and benefits, and the availability of alternative treatments before making a decision about a course of treatment. The requirement for informed consent is rooted in the concepts of an individual's right to bodily integrity and respect for patient autonomy").

benefits raised by this technology. Patient *z* in the case example above is left wondering precisely how Cardio *y* works. And no satisfactory answer is forthcoming. This is the challenge stated in its simplest terms: it is difficult, perhaps impossible, to obtain adequately informed consent when an essential component of an intervention is deeply unexplainable in the manner described in the first chapter. How unexplainable models work is necessarily uncertain. And this uncertainty might obstruct consent by depriving patients of a full account of the nature of the intervention about which they are asked to decide. Two things could be said about this. First, it may be that patients are generally not overly concerned about how, on a technical level, the devices and tools used their care function. Most patients likely do not, for example, enquire extensively about the way medical imaging devices or blood testing equipment works. In many clinical settings, just knowing what a tool is for and that it has been vetted and approved by a competent authority probably suffices. It could be that patients are generally not concerned, then, about the unexplainable character of AI models used in their care, so long as the models in question are not inherently dangerous. Second, the capacity of clinicians to provide detailed technical information about even traditional medical devices is likely limited. Many of the most widely used clinical tools probably have some degree of functional opacity from the perspective of the clinicians using them in patient care. While a cardiologist could surely explain how cardiovascular MR imaging works, they probably cannot reasonably be expected to know on a deeply technical level what specific imaging devices do. Expecting a clinician to know and explain the internal computing of a medical AI device might be like expecting a cardiovascular specialist to explain the computing that powers an MR imaging system.

These factors might suggest that the impact of unexplainable medical AI on the clinical obligation to obtain informed consent could well be limited. Patients probably do not generally want the kinds of explanations that are precluded by medical AI's often unexplainable character. And to expect clinicians to provide such explanations could be to expect far more of them than is expected in similar circumstances. But there remains a sense in which unexplainable medical AI could yet

generate difficulty for the law of informed consent. Unexplainable medical AI differs notably from other categories of clinical technology insofar as they operate in ways that are not just complex, but that are impossible to meaningfully describe. Though a clinician may not intimately understand the computing used in MR imaging, such computing is at least *in principle* knowable. The clinician could consult the device manual or request information from the manufacturer and in doing so probably have a relatively firm sense of how the device works. But this kind of inquiry for deeply unexplainable medical AI would not be very fruitful. That unexplainable AI is particularly resistant to technical comprehension might suggest that informed consent in the sense of a patient having the capacity to fully understand the course of a proposed intervention, is conceptually diminished. This could prompt clinicians to be disinclined to use unexplainable medical AI models in patient care, fearing that patient consent could be insufficiently informed. I suspect that this worry will not on its own have a significant effect on medical AI's uptake, at least not in the medium-term. I suspect that most patients and clinicians will be satisfied that medical AI models have been reviewed for safety and efficacy by the appropriate authorities, that there is evidence that their use contributes to improved patient outcomes, and that they way they work is, at least in broad terms, generally understood. That some of the details are opaque likely is not unique in this context and the conceptual distinction that the lingering opacity is necessary and irremediable likely will not be material in most practical clinical contexts. This is nevertheless an issue worth keeping in view as unexplainable medical AI proliferates and, as it proliferates, becomes ever more complicated and unexplainable. In the next section, I turn to an issue I think may have a more immediate effect on clinical practice: fault.

2. *Determining fault*

Suppose that not everyone is enthusiastic that Cardio *y* is being used for cardiovascular imaging and diagnosis. Imagine that a small but vocal cohort of clinical specialists argue publicly that Cardio *y*'s fundamentally unexplainable nature might subject patients to an unnecessary and unjustifiable degree

of risk. However popular, Cardio *y* has not yet become a standard of care. In the wake of Patient *z*'s death, their estate argues that Dr. *x*'s decision to rely on Cardio *y* constitutes a fault giving rise to liability for malpractice. A reasonably prudent clinician, the estate argues, would not have used a wholly unexplainable medical device. That no one understands just how the device functions suggests that clinicians using it are operating irrationally. It can never be prudent, they say, to use a medical device that is fundamentally unknowable. Dr. *x* claims that their conduct was reasonable and prudent in the circumstances. Many cardiologists, after all, use Cardio *y* for exactly the purpose for which it was used in Patient *z*'s treatment. It has been approved by Health Canada, suggesting that it is if nothing else safe and effective. Though not in universal use, the system is thought by many of Dr. *x*'s peers to be one of the most promising tools in cardiovascular imaging. In this first part of the chapter, I suggest that it is at present unclear which of these positions is the more persuasive. Whether Dr. *x*'s conduct was, from the perspective of civil liability, fault is not obvious. And it is not obvious in a specific and unique respect: owing to the unexplainable character of the AI model used in Patient *z*'s care, a reviewing court faces questions about the appropriate conduct of a clinician using these new medical devices and whether they can responsibly be used at all. This part is organized in three steps. First, I introduce civil law fault in the medical malpractice setting. Second, I give an overview of the common law duty of care that clinicians have conventionally been thought to owe their patients. In both traditions, clinician conduct is modulated by similar legal fictions: the notion of a reasonable clinician. Third, I outline how unexplainable AI makes trouble for this concept, likely complicating how jurists assess civil law fault and the common law duty of care.

a. Civil law fault

Responsibility for clinical malpractice in Quebec's civil law derives essentially from the law of contracts.[139] Clinicians provide medical services according to a contract for services formed between

[139] See André T Mécs, "Medical Liability and the Burden of Proof" (1970) 1:16 MLJ 163 at 164 ("It is presently generally accepted that the legal relationship between a doctor and his patient is contractual"); See also *X v Mellen*, [1957] BR 389 at

them and their patient. This contractual arrangement is often agreed tacitly.[140] The medical contract imposes certain obligations on clinicians. Most important for our present purpose is the obligation to provide medical services according to standards generally recognized within the profession, acting as a prudent and competent medical practitioner.[141] This is an obligation of means. The medical contract does not demand that clinicians obtain specific results, but only that they provide their services within a permissible range and exercising an appropriate level of professional skill.[142] A clinician bound under this contractual regime might be subject to liability for malpractice according to the notion of civil fault. Unlike in the common law system, in which liability is apportioned quite differently under contractual and tort law regimes, the civilian fault conceptually unifies private law liability.[143] There is significant ongoing debate among civil law scholars about fault's conceptual foundations. Whether,

410 ("J'estime que l'économie de notre droit civil impose l'impérieux devoir de s'y soustraire et d'y resister, ne serait-ce qu'au motif que si le demandeur, sur le champ de la responsabilité médicale, a la liberté de méconnaître les obligations contractuelles et de n'invoquer que les obligations légales du médecin, pourquoi ce principe s'appliquerait-il pas à tout autre domaine de responsabilité?"); Philips-Nootens et al, *supra* note 120 at 2 ("Les relations juridiques entre le médecin et son patient peuvent naître d'un contrat ou relever, en l'absence de contrat, d'obligations imposées par le législateur. Si la plupart des obligations se retrouvent dans ces deux types de situations, chacune d'elles comporte néanmoins des particularités quant à un éventuel recours en responsabilité").

[140] See *X v Mellen*, *supra* at 408–409 ("Dès que le patient pénètre dans le cabinet de consultation du médecin, prend naissance entre celui-ci et le malade, par lui-même ou pour lui-même, un contrat de soins professionnels"); Philips-Nootens et al, *supra* note 120 at 10 ("'Le contrat se forme par le seul échange de consentement entre des personnes capables de contracter' à moins que la loi n'exige en outre des formalités particulières, ce qui n'est pas le cas du contrat médical").

[141] Philips-Nootens et al, *supra* note 120 at 55 ("Elle comporte celle de se conformer à des standards généralement reconnus dans la profession. Le critère applicable demeure celui du praticien normalement prudent et compétent").

[142] Philips-Nootens et al, *ibid* at 55 ("Hormis l'obligation de résultat que l'on peut retrouver à propos de l'utilisation des appareils ou le respect du secret, de façon constante, la doctrine et la jurisprudence ont défini l'obligation du médecin comme une obligation de moyens… Pour reprendre l'expression du professeur Crépeau, il faut distinguer entre l'art médical et la technique médicale. Le médecin a le devoir d'agir avec diligence et habileté, « d'exercer sa profession selon les normes médicales actuelles les plus élevées possibles », mais on ne peut attendre de lui, de ce seul fait, la guérison du patient : il y a en jeu trop d'éléments qu'il ne peut contrôler et qui tiennent à l'organisme du malade, au comportement de celui-ci, à l'évolution de la maladie, à l'avancement de la science").

[143] Jean-Louis Baudouin, Patrice Deslauriers & Benoît Moore, *La responsabilité civile, Volume 1: Principes généraux*, 9th edition (Montreal: Éditions Yvons Blais, 2020) at 1-46 ("Conceptuellement, les différences fondamentales entre responsabilité contractuelle et responsabilité extracontractuelle s'estompent donc, puisque toutes deux entraînent une obligation de réparation ayant pour origine le manquement à une obligation préexistante soit d'ordre conventionnel (responsabilité contractuelle), soit d'ordre extracontractuel (responsabilité légale). Une seule différence sépare les deux. Alors qu'en règle générale la seconde résulte du manquement à une obligation de ne pas faire, permanente et légale, et résulte d'un fait juridique, la première peut résulter d'une contravention à une obligation de faire ou de ne pas faire, est temporaire et prend sa source dans un acte juridique").

for example, fault should be thought to convey specific moral content—whether the commission of a fault ought to be conceived as a moral wrong—remains a contentious and unsettled question.[144] Apart from this debate in fault theory, contemporary civilian jurists agree that fault refers to the breach of an obligation: *le manquement à un devoir*.[145] This is a rule found in Article 1458 of the *Civil Code*, which articulates both an obligation to honour contractual undertakings and to repair injuries caused by breach of that obligation.[146] Failing to meet a contractually-imposed obligation, as in the case of medical malpractice, constitutes a fault for which liability may be imposed.

In clarifying precisely what this fault standard demands, civilians sometimes distinguish between fault assessed *in concreto* and fault assessed against an objective, *in abstracto* measure. The *in concreto* standard frames fault according to an individual's prior conduct.[147] A clinician would fail to meet the obligation imposed by contract only to the degree that they deviate from their customary or normal practice. This vision is widely rejected, for it faces the obvious objection that an habitually

[144] See Baudouin, Deslauriers & Moore (volume 1), *ibid* at 1-161 ("Le droit québécois, à la différence des droits d'inspiration germanique, n'a pas retenu le concept de l'illicéité. Selon une auteure la notion d'illicéité a toutefois été introduite dans le droit commun de la responsabilité à l'article 1457, al. 1 C.c. Pour elle, la notion de faute, présente à l'alinéa, est la conjonction d'un élément matériel (l'acte illicite) et subjectif (l'imputation de cet acte à une personne ayant la capacité de discernement"); Marième Lacroix, "Le fait générateur de responsabilité civile extracontractuelle personnelle: continuum de l'illicéité à la faute simple, au regard de l'article 1457 C.c.Q." (2012) 46:1 Revue juridique Thémis 25 at 34 ("En vertu de l'alinéa premier de l'article 1457 C.c.Q., le législateur définit l'illicéité comme un manquement au devoir de respecter les règles de conduite. L'illicéité est ici purement matérielle et objective: il s'agit de la contravention à un devoir de bonne conduite, à une norme de civilité").

[145] Philips-Nootens et al, *supra* note 120 at 49 ("Les auteurs soulignent la difficulté que comporte toute tentative de définition de la faute, en raison des nombreux éléments qui peuvent entrer en jeu. Une constante se dégage cependant: le manquement à un devoir"); Baudouin, Deslauriers & Moore (volume 1), *supra* note 147 at 1-163 ("D'une façon générale, la plupart des définitions données par la doctrine se regroupent autour de deux idées maîtresses: le manquement à un devoir préexistant et la violation d'une norme de conduit").

[146] Art 1458 CCQ ("Every person has a duty to honour his contractual undertakings. Where he fails in this duty, he is liable for any bodily, moral or material injury he causes to the other contracting party and is bound to make reparation for the injury; neither he nor the other party may in such a case avoid the rules governing contractual liability by opting for rules that would be more favourable to them").

[147] Baudouin, Deslauriers & Moore (volume 1), *supra* note 147 at 1-194 ("L'appréciation *in concreto* consiste à mettre en parallèle la conduite habituelle de l'auteur du préjudice et celle qu'on lui reproche d'avoir eue au moment où il a causé le dommage. Dans cette perspective, il y a faute si son comportement n'est pas conforme à celui qu'il a l'habitude d'avoir").

imprudent person could escape liability just by being consistent in their carelessness.[148] Approaching fault *in abstracto* prompts a reviewing court to consider how an individual's conduct measures up against the conduct of an abstract reasonable person acting under similar conditions.[149] Taking this view, fault would constitute "the gap between an individual's actual conduct and that of a reasonable, prudent, and diligent person," the archaic *bon père de famille*.[150] In medical malpractice, a clinician's *in abstracto* fault consists of failing to meet the contractually-imposed standard of a prudent and competent medical practitioner. Clinicians are bound to practice their profession with diligence and skill, according to current medical practices, and in alignment with the customary conduct of their fellow professionals.[151] I specified above that this is an obligation of means. Clinicians are generally not obliged to achieve defined outcomes, in contrast with obligations of result or warranty.[152] One consequence of this characterization is that errors in clinical judgment are on their own insufficient to prove malpractice.[153] In effect, this means that courts assessing fault generally do not focus on the outcome of a particular act or omission, but rather on whether the comportment of a defendant

[148] Philips-Nootens et al, *supra* note 120 at 61 ("On perçoit tout de suite l'issue d'une telle démarche: le médecin habituellement prudent serait condamné pour le moindre écart de conduite, tandis que celui qui est généralement peu consciencieux devrait commettre une faute grossière pour être condamné").

[149] Baudouin, Deslauriers & Moore (volume 1), *supra* note 147 at 1-195.

[150] Baudouin, Deslauriers & Moore (volume 1), *ibid* at 1-195 ("L'appréciation *in abstracto* retenue par le droit civil permet, au contraire, de répondre à ces objections. La faute civile extracontractuelle est constituée par l'écart séparant le comportement de l'agent de celui du type abstrait et objectif de la personne raisonnable, prudente et diligente, du bon citoyen (du «bon père de famille», disait-on auparavant"); Alexandra Popovici, "Le bon père de famille" in Générosa Bras Miranda et Benoit Moore, eds, *Mélanges Adrian Popovici. Les couleurs du droit* (Montréal, Éditions Thémis, 2010) 125–141.

[151] See Paul-André Crépeau, *L'intensité de l'obligation juridique, ou, Des obligations de diligence, de résultat et de garantie* (Montreal: Éditions Yvons Blais, 1989) at 51.

[152] Philips-Nootens et al, *ibid* at 55 ("Hormis l'obligation de résultat que l'on peut retrouver à propos de l'utilisation des appareils ou le respect du secret, de façon constante, la doctrine et la jurisprudence ont défini l'obligation du médecin comme une obligation de moyens"); Jean-Louis Baudouin, Patrice Deslauriers & Benoît Moore, *La responsabilité civile, Volume 2: Responsabilité professionelle*, 9th edition (Montreal: Éditions Yvons Blais, 2020) at 2-34 ("Le médecin, comme tout professionnel, est tenu en principe à l'endroit du patient à une obligation de moyens. En d'autres termes, il doit, dans le diagnostic et le traitement, se comporter comme un médecin raisonnablement prudent et diligent placé dans les mêmes circonstances") [Baudouin, Deslauriers & Moore (volume 2)].

[153] See *Lapointe c Hôpital Le Gardeur*, [1992] 1 SCR 351 at 363, 90 DLR (4th) 7 ("Les professionnels de la santé ne devraient pas être tenus responsables de simples erreurs de jugement, qui sont distinctes de la faute professionnelle").

physician is conduct that a reasonable and prudent professional would have taken in the same circumstance.[154] Justice Gonthier in the Supreme Court's *St. Jean v. Mercier* decision puts it this way:

> To ask, as the principal question in the general inquiry, whether a specific positive act or an instance of omission constitutes a fault is to collapse the inquiry and may confuse the issue. What must be asked is whether that act or omission would be acceptable behaviour for a reasonably prudent and diligent professional in the same circumstances. The erroneous approach runs the risk of focussing on the result rather than the means. Professionals have an obligation of means, not an obligation of result.[155]

To summarize, fault applied to clinicians has roughly the following essential contours. Fault is the violation of an obligation. In the medical context, clinicians have an obligation to practice their profession according to the standard of a reasonable and prudent clinical professional. This is an obligation not to achieve a specific outcome, but to exercise appropriate diligence in the care of patients. In the following part, I give an overview of the concept of the duty of care as it applies to clinicians in the common law. As I suggest there, fault in Quebec's civil law tradition shares a great deal with the common law's breach of the standard of care.

b. Common law fault

Much as the concept of fault in civil law is a foundational element in civil liability, the duty of care is conventionally understood to be a necessary condition for liability in the common law of torts.[156] Medical malpractice in Canadian common law is addressed as a species of the tort of negligence.[157] With its proximate origins in the famous *Donoghue v. Stevenson* case decided by the House of Lords in 1932, negligence is founded on a broad duty not to cause injury to one's neighbours.[158] As a

[154] Philips-Nootens et al, *supra* note 120 at 55.

[155] *St–Jean v Mercier*, 2002 SCC 15 at para 53, [2002] 1 SCR 491.

[156] See e.g. Donal Nolan, "Deconstructing the Duty of Care" (2013) 129 LQ Review 559 at 559 ("The existence of what is termed a 'duty of care' is generally regarded as a fundamental building block of the common law of negligence, a 'core ingredient' or a 'foundational element' of the cause of action").

[157] Hardcastle, *supra* note 116 at 306.

[158] *Donoghue v Stevenson*, [1932] UKHL 100 at 580, [1932] All ER Rep 1 ("The rule that you are to love your neighbour becomes in law, you must not injure your neighbour; and the lawyer's question, Who is my neighbour? receives a restricted

condition for establishing a claim in civil liability, plaintiffs are required to demonstrate that a duty of the kind described in *Donoghue* was owed to them by the defendant.[159] Multiple such duties of care have been recognized by Canadian courts, applying to a wide variety of circumstances with varying degrees of specificity.[160] The controlling Supreme Court opinion for establishing a duty of care, *Cooper v. Hobart*, clarifies a two-part test first enunciated by the House of Lords in *Anns v. Merton London Borough Council*.[161] In the first part of the *Cooper* test, a court considers two factors: whether the injury complained by the plaintiff was a reasonably foreseeable consequence of the defendant's action and whether the plaintiff and defendant were in a sufficiently proximate relationship to give rise to a *prima facie* duty of care.[162] In the second part of the test, a court assesses whether there exist any residual policy considerations, specifically with respect to the effect of recognizing the existence of a duty of care between the parties, that would mitigate in favour of negating a *prima facie* duty established in the test's first step.[163] Absent any such consideration, a reviewing court will conclude that a duty of care existed between the parties. There is rarely any

---

reply. You must take reasonable care to avoid acts or omissions which you can reasonably foresee would be likely to injure your neighbour").

[159] See Ernest J Weinrib, *Corrective Justice* (Oxford: Oxford University Press, 2012) at 44 ("The signal achievement of negligence law in the twentieth century was to develop the concepts of negligence analysis in a way that coherently links the unreasonable risk to the harm suffered. Duty and proximate cause are crucial components in this linkage. These concepts connect fault and injury by describing the wrongful risk in terms of the range of the potential victims and consequences through which the risk is to be understood as wrongful. Duty connects the defendant as a wrongdoer to the plaintiff as a member of the class of persons wrongfully put at risk").

[160] Kristen Thomasen, "AI and Tort Law" in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 103–122 at 111.

[161] *Cooper v Hobart*, 2001 SCC 79 at para 30, [2001] 3 SCR 537.

[162] *Cooper v Hobart*, *ibid* at paras 30–31 ("If foreseeability and proximity are established at the first stage, a prima facie duty of care arises").

[163] *Cooper v Hobart*, *ibid* at paras 37–38 ("As the majority of this Court held in Norsk, at p. 1155, residual policy considerations fall to be considered here. These are not concerned with the relationship between the parties, but with the effect of recognizing a duty of care on other legal obligations, the legal system and society more generally. Does the law already provide a remedy? Would recognition of the duty of care create the spectre of unlimited liability to an unlimited class? Are there other reasons of broad policy that suggest that the duty of care should not be recognized?").

meaningful dispute that a duty of care exists between clinicians and their patients.[164] Clinical duties of care are well established in Canadian law,[165] and are usually only in dispute where there is unclarity about whether a clinician-patient relationship was established or terminated,[166] or where an alleged injury affects a third person not in direct relationship with the defendant clinician.[167]

That a duty of care straightforwardly exists between clinical professionals and their patients does not reveal much of what the duty requires. As conventionally expressed in *Donoghue*, the common law duty of care demands that individuals conduct themselves such that they avoid causing injury to those with whom they are in proximate relation. Common law courts have traditionally measured breach of the duty of care owed by clinicians according to a standard of a "normal, prudent practitioner of the same experience and standing."[168] This approach suggests that specialist clinicians are generally bound to a more exacting duty of care than their generalist counterparts and that a clinician's relative degree of experience is a factor in articulating the standard of care to which

[164] Hardcastle, *supra* note 116 at 307–308.

[165] See Gerald B Robertson & Ellen I Picard, *Legal Liability of Doctors and Hospitals in Canada*, 5th Edition (Toronto: Thomson Carswell, 2017) at 269.

[166] There is typically little question that a clinician-patient relationship has been established. In some cases, subject to provisions of discrimination law, a clinician may be permitted not to accept a patient. Where a clinician permissibly declines to act in a professional capacity for an individual, no clinician-patient relationship is established and, in consequence, no duty of care extends between the parties. See Hardcastle, *supra* note 116 at 308; *HH v RG*, Health Professionals Appeal & Review Board, ON (2013), 11–CRV–0178 at paras 38–44 ("In the present case, the question has been raised as to whether [the Respondent's] practice of not treating patients who smoke is discriminatory on the basis of disability, and if so, whether he met his obligation to accommodate [the patient] up to the point of undue hardship").

[167] Genetic medicine, for example, is especially likely to produce clinical considerations that implicate not only individual patients, but the relatives of patients as well. In the event genetic test results indicate a that serious condition may affect relatives of a patient, it may be that a duty of care on the part of a treating physician could be established with respect to family members with whom the clinician has no formal, legally cognizable professional relationship. These cases are outliers; in the overwhelming majority of clinical interactions, a duty of care derived from a clinician-patient relationship will be obvious and uncontested. Mark A Rothstein, "Reconsidering the duty to warn genetically at-risk relative" (2018) 20 Genetics in Medicine 285 at 285 ("The theory that physicians are legally required to warn their patients' relatives when the patients fail to do so, even over the objection of their patients, raises serious concerns about professional responsibility and possible conflicts with federal health privacy law"); Adrian Thorogood, Alexander Bernier, Ma'n H Zawati & Bartha Maria Knoppers, "A Legal Duty of Genetic Recontact in Canada" (2019) 40:2 Health Law in Canada 58 at 59 ("There is also the possibility for novel legal claims, including failures to warn relatives of genetic risks (even where the information is confidential to the patient"); *Watters v White*, 2012 QCCA 257, JE 2012-473.

[168] *Crits v Sylvester*, [1956] OR 132 at 13, 1 DLR (2d) 502 (Ont. CA).

medical professionals are bound.[169] As in the civil law approach to medical fault, clinicians are not expected to be free of error in performing their profession. In *Wilson v. Swanson*, the Supreme Court clarifies that "the honest and intelligent exercise of judgement has long been recognized as satisfying the professional standard," even in the presence of medical error.[170] What is important is that the clinician comports themselves as would a similarly situated, reasonably prudent clinician in the same circumstance. Framed in this way, the common law duty of care that clinicians owe their patients has much in common with the civil law concept of professional fault in medicine. In both traditions, assessing whether a clinician's action constitutes fault turns on the degree to which the clinician behaved as a reasonable and prudent professional would have behaved.

c. An unexplainably (un)reasonable clinician

In the sections above, I outlined how fault is characterized in civil and common law malpractice regimes. Both traditions are of effectively the same mind in setting out a standard of fault responsive to professional conduct that does not meet the standard of a reasonable and prudent clinician in like circumstances. Unexplainable AI makes it potentially more difficult to conceptualize what this standard precisely requires of clinicians. Whether a reasonable and prudent clinician ought, for example, to use unexplainable AI merely as a supplement to professional judgement or as a reference afforded a high degree of deference, does not admit of an obvious answer. Some of the difficulty for medical practice is probably highly abstract. It might be difficult to imagine, for example, how deeply unexplainable medical AI could ever reasonably and prudently be applied in patient care if no one, not even a model's initial programmer, fully understands how it works. Assuming that the reasonable clinician construct presupposes some minimal understanding of the way clinical interventions function, then the application of an unexplainable model might in some sense operate as an inherently

[169] Hardcastle, *supra* note 116 at 311; See also Thomasen, *supra* note 148 at 114–115.

[170] *Wilson v Swanson*, [1956] SCR 804 at 812, 5 DLR (2d) 113.

imprudent venture.[171] We might intuitively think, in other words, that it would not generally be reasonable for physicians to rely on medical implements that operate in some entirely incomprehensible way. But it is difficult to see how this position would align with the law's prevailing conceptions of medical fault, attached as they are to the referent of standard professional practice. And anyway, as I describe in the third chapter below, this vision does not much accord with the way clinical care is delivered. Much of medical science relies on fuzzy reasoning and on judgement that is challenging or impossible to explain *ex post*.[172] But there is nevertheless a kind of intuitive uneasiness that might be attached to the clinical use of deeply unexplainable AI. While no one fully knows how acetaminophen works, for example,[173] it is at least theoretically in our capacity, with contemporary biological and chemical knowledge, to find out. Deeply unexplainable AI might reasonably feel quite different. And while I do not think this attitude is ultimately persuasive, it gestures at the complexity of the challenge for assessing the reasonable clinician standard.

More tangibly, unexplainable medical AI might prompt debate about the proper disposition of the reasonable clinician toward patient care. Assuming medical AI will at some not-so-distant future moment systematically and consistently surpass human clinical judgement in terms of clinical

[171] See Juan Manuel Durán & Karin Rolanda Jongsma, "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI" (2020) 47 Journal of Medical Ethics 329 at 329 ("The epistemological opacity that characterises black box algorithms seems to be in conflict with much of the discursive practice of giving and asking for reasons to believe in the results of an algorithm, which are at the basis of ascription of moral responsibility. Concerns relate to problems of accountability and transparency with the use of black box algorithms, (hidden) discrimination and bias emerging from opaque algorithms, and the raising of uncertain outcomes that potentially undermine the epistemic authority of experts using black box algorithms"); Rudin & Radin, *supra* note 85 at 3; Rudin, *supra* note 94 at 206–207.

[172] See Gunver S Kienle & Helmut Kiene, "Clinical judgement and the medical profession" (2011) 17 Journal of Evaluation in Clinical Practice 621 at 621 ("Initially, the clinically skilled and scientifically competent doctors and their judgements were the main impetus for treatment decision, therapy assessment and medical progress. With the rise of modern research methodology, however, the fallacious aspects of clinical judgement were increasingly emphasized. It was presumed that personal judgement would be unable to go beyond a simple *post hoc ergo propter hoc*, and could at best accomplish something like simple, intuitive, low-quality correlational statistics").

[173] Ghassemi et al, *supra* note 93 at 748 ("Despite competing explanations for how acetaminophen works, we know that it is a safe and effective pain medication because it has been extensively validated in numerous randomised controlled trials (RCTs). RCTs have historically been the gold-standard way to evaluate medical interventions, and it should be no different for AI systems").

accuracy, then it may be that the proper standard is one of general deference to even unexplainable models.[174] This is a position taken by jurists adopting the view that non-deference to a provably more accurate mode of diagnosis and care allocation would effectively deprive patients of the highest level of available care. Portrayed on these terms, non-deference to a discernibly more accurate mode of medical decision-making might constitute malpractice. But no such clinical standard has yet, so far as I know, been enforced anywhere on the planet. And if it were, this kind of relationship to AI would constitute a major shift in some of our most compelling juristic and ethical intuitions about medical practice.[175] To turn over significant decision-making authority to computer models under the direction of medical malpractice law would reasonably make many of us uncomfortable. This attitude would also directly conflict with perspectives dominant in AI ethics guidance, particularly the view that unexplainable AI models should be supplemented by the input or oversight of a human decision-maker. This is the view taken by advocates of 'human-in-the-loop' AI oversight. The *Montreal Declaration for a Responsible Development of AI*, for example, warns against using AI to replace human beings in duties that "require quality human relationships," which likely quintessentially includes the practice of medicine.[176] And while human-in-the-loop AI oversight is, I think, severely

[174] See 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 at 61–62 ("Thus, a physician, hospital, or insurer relying on an ML diagnosis will, at least initially, be held to no higher standard than that of the ordinary physician. Once ML itself becomes the standard of care, ML will raise the bar. But even though a higher level of accuracy will now be the standard, the malpractice exposure of ML-users will actually shrink because by relying on ML they will be complying with the professional standard; at that point, reliance on human diagnosticians will become the risky legal strategy both for failing to use an increasingly common technology of which they should have been aware and because (by hypothesis) the risk of error is in fact greater").

[175] See Michael Lang, Alexander Bernier & Bartha Maria Knoppers, "AI in Cardiovascular Imaging: 'Unexplainable' Legal and Ethical Challenges?" (2021) Canadian Journal of Cardiology [in press, doi.org/10.1016/j.cjca.2021.10.009] ("As ML systems become increasingly proficient, legal and ethical pressure might prompt physicians to delegate important care decisions to automated processes. While this might improve care under optimal conditions, it might also produce practice decisions that human doctors are unable to understand. This would dramatically change the practice of medicine while also undercutting some of our most dominant ethical intuitions about automated decision-making").

[176] See e.g. Marc-Antoine Dilhac, Christophe Abrassart, Nathalie Voarino et al, "Montreal Declaration for a Responsible Development of Artificial Intelligence" (Montreal: Université de Montréal, 2018) at principle 4 ("[AI systems] should not be implemented to replace people in duties that require quality human relationships, but should be developed to facilitate these relationships") [Montreal Declaration].

conceptually limited, it is certainly true that a standard of care that would lean even partially toward AI-deference would constitute a major shift in the effect of malpractice law on clinical behaviour.

Much more immediately pressing, though, are questions surrounding how clinicians ought to use presently available unexplainable models in their practice and how coordination between AI and human-mediated decision-making ought to be evaluated. Put more directly, the standard of care applicable to the use of unexplainable AI is not yet clearly defined.[177] Because courts assess clinical fault according to a field's general practice, what a clinician should do is shaped in large measure by the customary practices of the broader clinician community.[178] In measuring a clinician's conduct against professional custom, courts might rely on expert opinion evidence[179] or professional practice guidance prepared by medical colleges or other professional associations.[180] Clinicians in medical practice will naturally rely significantly on practice guidance in structuring their approach to patient care, particularly for newly emerging medical innovations.[181] In the case of medical AI, whether

[177] See W Nicholson Price III, Sara Gerke & I Glenn Cohen, "Potential Liability for Physicians Using Artificial Intelligence" (2019) 322:18 JAMA 1765 at 1765 ("In general, to avoid medical malpractice liability, physicians must provide care at the level of a competent physician within the same specialty, taking into account available resources. The situation becomes more com- plicated when an AI algorithmic recommendation becomes involved. In part because AI is so new to clinical practice, there is essentially no case law on liability involving medical AI").

[178] Sullivan & Schweikart, *supra* note 14 at 161 ("In judicial determinations, a physician's actions are judged not against those of a reasonable *man*, but rather against those of a reasonable physician—with the same knowledge, skills, and expertise—under like circumstances. However, courts do not purport to possess the knowledge necessary to determine sound medical judgment. Thus, expert testimony of qualified physicians is required to establish the standard of care or what is 'reasonable to expect of a professional given the state of medical knowledge at the time of the treatment in issue.' Given the nature of medical practice, custom is largely dispositive").

[179] Tobia et al, *supra* note 13 at 17 ("This customary standard is normally supported by expert witness testimony, clarifying the local or national practice. Often, jurors evaluate what the normal or average physician would do in light of conflicting testimony from dueling medical experts").

[180] Nancy MP King, "The Reasonable Patient and the Healer" (2015) 50 Wake Forest Law Review 343 at 346 ("Similarly, in professional negligence, which includes medical malpractice, it is the actions of the professional that come under scrutiny. The standard against which the actor is judged is that of the relevant professional under the circumstances-measured by expert testimony, professional guidance, etc.").

[181] Paul W Armstrong, "Do Guidelines Influence Practice?" (2003) 89 Heart 349 at 352 ("What is the evidence that guidelines can provide a meaningful impact on medical practice? Grimshaw and Russell identified 59 published evaluations of clinical guidelines that met defined criteria for scientific rigour and concluded that explicit guidelines could improve clinical practice…it is evident that the development strategy, method of dissemination of the guidelines, how they are implemented and what process of evaluation exists are key to the likelihood of them being effective").

unexplainable or not, not much in the way of formal professional clinical guidance yet exists, and much of it applies only in narrow practice bands or specialties. An early professional position paper, published by an AI working group of the Canadian Association of Cardiologists, provides recommendations for the Association, promotes engagement with regulatory agencies on ethical and legal issues surrounding the clinical adoption of AI, and supports common standards for AI testing and validation.[182] But it does not set out in detail how radiologists ought to use (or not use) AI models in patient care. Likewise, the Royal College of Physicians and Surgeons in its recommendations on the clinical implementation of AI, suggests that present legal norms are probably insufficient to address questions surrounding the liability of clinicians for injuries associated with AI-mediated decision-making,[183] but does not set out rules or firm practice guidelines. It might be unsurprising that this should be so, for medical AI, with all of its theorized promise, has not yet been widely adopted in patient care. But the character of present professional guidance, I think, underscores the present uncertainty facing clinicians with respect to appropriate standards of practice for AI use.

There may be good reason to suspect that it will be especially challenging to define practice standards for unexplainable AI. Models that do not admit of explanation might be, as I suggested above, particularly challenging to review. And when unexplainable models make mistakes, it will often be impossible to explain how it happened, making the sources of mistake significantly more difficult to address. One well-documented source of potential error, for example, might be unintentionally biased decision-making. To the degree models are trained on demographically

[182] An Tang et al, "Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology" (2018) 69 Canadian Association of Radiologists Journal 120 at 125 ("The CAR must engage with regulatory agencies on the ethical and medico-legal issues concerning AI …The CAR should support and develop common standards for validation and testing of AI tools, emphasizing stability of performance over varying settings, equipment, and protocols to certification for clinical use").

[183] Royal College of Physicians and Surgeons, *Task Force Report on Artificial Intelligence and Emerging Digital Technologies* (Ottawa: Council Task Force on Artificial Intelligence and Emerging Digital Technologies, 2020) at 36 ("New regulatory frameworks are required that emphasize timely implementation of legal and ethical considerations, such as explainability and transparency, prevention of bias and discrimination, data-related matters, privacy and security, and liability and accountability").

unrepresentative patient datasets, for example, clinical outputs may be inadequately responsive to certain groups.[184] Training data collected from disproportionately white, wealthy, and able-bodied populations is likely to produce AI models that are disproportionately responsive to these groups while underserving others. It is now well understood that this stands as one of the predominant risks associated with the use of AI models in health and other sensitive domain. For AI models that are unexplainable, the challenge is greater still. We might know that a model's outputs are biased in one way or another but be entirely incapable of knowing why or how to fix it. Each of these factors might reasonably cause reticence on the part of medical colleges and professional bodies in recommending that these kinds of models feature in clinical practice. Of course, challenges in defining clinician obligations are, in one sense, easily remedied. What the law expects from medical professionals will be settled as the professionals themselves define emerging standards of practice and reconceptualize their clinical function. Malpractice law has adjusted repeatedly in response to novel medical innovation throughout its history and will certainly do so again in response to unexplainable AI. But because unexplainable AI is uniquely singular as a medley of disruptive promise and functional opacity, the uncertainty it produces might have an especially prominent salience. I will have more to say about this below and in the third chapter. In the next part, I consider the effect of unexplainable medical AI on another foundational element of medical malpractice law: causation.

3. *Assessing Causation*

Imagine that Dr. *x* definitively owes Patient *z* an obligation to provide medical services according to the standard of a reasonable and prudent clinician and that this obligation includes a directive to cross-

[184] See Ravi B Parikh, Stephanie Teeple & Amol S Navathe, "Addressing Bias in Artificial Intelligence in Health Care" (2019) 322:24 JAMA 2377 at 2377 ("Recent scrutiny of artificial intelligence (AI)–based facial recognition software has renewed concerns about the unintended effects of AI on social bias and inequity. Academic and government officials have raised concerns over racial and gender bias in several AI-based technologies, including internet search engines and algorithms to predict risk of criminal behavior"); Sandeep Reddy, Sonia Allan, Simon Coghlan & Paul Cooper, "A governance model for the application of AI in health care" (2020) 27:3 Journal of the American Medical Informatics Association 491 at 492.

reference the outputs of unexplainable medical AI models with Dr. *x*'s own clinical expertise and experience. In other words, imagine that the courts have recognized an obligation on the part of clinicians using unexplainable AI models to confirm machine-generated decisional outputs. Now imagine that, as above, Dr. *x* diagnoses Patient *z* with a rare and complex ventricular abnormality and that this diagnosis was partially informed by a report generated by Cardio *y*. But unlike in our earlier scenarios, Dr. *x* this time independently considers alternate possible diagnoses. Ruling them out, Patient *z* is again misdiagnosed and dies a short while longer. Patient *z*'s estate again commences a malpractice claim against Dr. *x*. Though Patient *z*'s estate succeeds, by assumption, in asserting fault, they might yet face significant difficulty making out the balance of their claim. Proving that Dr. *x*'s fault was the cause of Patient *z*'s injury might not be straightforward. In making the case for this view, I approach the concept of causation, as I did with fault above, through distinct civil and common law lenses. In the common law tradition, an important distinction is drawn between causation in fact and causation in law. I will address each of them in turn. No such firm distinction is made in Quebec's civil law, though there is much conceptual overlap in the ways each tradition assesses whether a defendant produced a plaintiff's injury. Causation has long been thought to be, equally in the common law as in civil, the "most difficult task in medical malpractice litigation."[185] I suggest it might be made appreciably more difficult where an alleged fault occurs in the use of unexplainable AI. Whether Dr. *x* caused Patient *z*'s injury, may not be readily apparent.

a. Common law causation

Causation in the common law of civil liability is composed of two subsidiary concepts: causation in fact and causation in law. The logically primary of them is factual causation, which serves as a

[185] Dieter Geisen, *International Medical Malpractice Law: A Comparative Law Study of Civil Liability Arising from Medical Care* (Tübingen: JCB Mohr, 1988) at 163; See also Lara Khoury, *Uncertain Causation in Medical Liability* (London: Bloomsbury Publishing, 2006) at 13–14.

condition for the legal attribution of causation to a defendant's fault.[186] Factual causation effectively asks whether the plaintiff's injury would have occurred absent the defendant's action.[187] An analysis of factual causation usually proceeds *ex hypothesis*, with an imagining of the world as it would have been absent the conduct alleged to be the source of the plaintiff's injury.[188] This way of thinking about causation in fact is typically described by jurists and scholars as the 'but for' approach.[189] It notably implies, as the famous English High Court case *Barnett v. Chelsea Hospital* illustrates, that defendants will generally not be subject to liability for patient injuries that would have occurred in any case, or which had already occurred before the defendant's fault.[190] In *Barnett*, three night watchmen attended a local hospital after drinking arsenic-tainted tea.[191] They were sent home without seeing a doctor and one of the men died a short time later. His spouse sued on behalf of the estate, arguing that the hospital had breached a duty of care owed to the three men by not admitting them for treatment. The High Court agreed but found that the plaintiff's death would have occurred in any

[186] See e.g. Bernard Dickens, "Medical Negligence" in Jocelyn Grant Downie, Timothy A Caulfield & Colleen M Flood, eds, *Canadian Health Law and Policy*, 4th Edition (Markham: LexisNexis Canada, 2011) 113–151 at 136.

[187] See *Snell*, *supra* note 122 at 302 ("Both the trial judge and the Court of Appeal relied on McGhee, which (subject to its re-interpretation in the House of Lords in Wilsher) purports to depart from traditional principles in the law of torts that the plaintiff must prove on a balance of probabilities that, but for the tortious conduct of the defendant, the plaintiff would not have sustained the injury complained of"); See also Ernest Weinrib, "A Step Forward in Factual Causation" (1975) 38 Modern Law Review 518 at 522 ("But instead of probing the relationship between the conduct and the injury the 'but for' test considers the injury in isolation from the conduct. It does not assess the defendant's conduct directly but rather the situation in the absence of that conduct, and this situation is ex hypothesis non-existent. From the examination of the primary question of what happened, we slide into the question of what did not happen or rather what would have happened if what had happened had not happened") [Weinrib 1975].

[188] See Weinrib 1975, *supra* at 522.

[189] See Khoury, *supra* note 205 at 18 ("Factual causative inquiry is most frequently carried out with the assistance of the but-for test, which has met with near universal acceptance as a tool for achieving the determination").

[190] Khoury, *supra* note 205 at 19.

[191] *Barnett v Chelsea and Kensington Hospital Management Committee*, [1968] 1 All ER 1068 at 1068, [1969] 1 QB 428 ("At about 5 a.m. on Jan. 1, 1966, three night watchmen drank some tea. Soon afterwards, all three men started vomiting. At about 8 am the men walked to the casualty department of defendants' hospital, which was open. One of them, deceased, when he was in the room in the hospital, lay on some armless chairs. He appeared ill. Another of the men told the nurse that they had been vomiting after drinking tea. The nurse telephoned the casualty officer, a doctor, to tell him of the men's complaint. The casualty officer, who was himself unwell, did not see them, but said that they should go home and call in their own doctors. The men went away, and deceased died some hours later from what was found to be arsenical poisoning").

event, for no reliable method of treatment could have been administered in sufficient time to save him.[192] So, negligence did not, as factual matter, cause the plaintiff's injury. He would have died even if the physician had acted as they ought to have done. Canada's Supreme Court affirmed *Barnett* in *Clements v. Clements*, finding that an injured plaintiff may recover only on proving that the defendant's intersession was necessary in bringing about their injury.[193] Before *Clements*, courts regularly deviated from the empirically strict 'but for' standard, permitting recovery, for example, on a material contribution theory of causation according to which defendants could be liable for negligent actions or omissions that merely contributed to a plaintiff's injury to a degree greater than *de minimus*.[194] *Clements* specifies that causation in fact does not require "scientific evidence of the precise contribution of the defendant's negligence" to an injury.[195] Reviewing courts are allowed some flexibility and might draw "common sense" inferences in determining factual cause.[196]

Causation in fact is by itself inadequate for recovery. Plaintiffs must also prove that injuries factually caused by a defendant's fault are legally cognizable. Legal causation, or causation in law,

---

[192] *Barnett*, *ibid* at 1074 ("Dr. Goulding said this in the course of his evidence: … 'I see no reasonable prospect of the deceased being given [the curative antidote] before the time at which he died.' … I regard that evidence as very moderate, and that might be a true assessment of the situation to say that there was no chance of [the curative antidote] being administered before the death of the deceased").

[193] *Clements v Clements*, 2012 SCC 32 at para 8, [2012] 2 SCR 181 ("The test for showing causation is the "but for" test. The plaintiff must show on a balance of probabilities that "but for" the defendant's negligent act, the injury would not have occurred. Inherent in the phrase "but for" is the requirement that the defendant's negligence was necessary to bring about the injury — in other words that the injury would not have occurred without the defendant's negligence. This is a factual inquiry").

[194] Khoury, *supra* note 205 at 22 ("The material contribution test thus allows for a departure from the traditional but-for test of causation and flows in fact from the recognition that this test is sometimes unworkable or, more precisely, leads to results considered unfair. The more liberal material contribution test considers it sufficient to show that the defendant's negligence materially contributed to producing the damage, even though his act alone was not sufficient to create it. Materiality is a question of degree and is met by any contribution which does not fall within the exception *de minimus non curat lex*"); Hardcastle, *supra* note 116 at 316.

[195] *Clements*, *supra* note 181 at para 9; See also Hardcastle, *supra* note 116 at 316.

[196] *Clements*, *supra* at para 9; Khoury, *supra* note 205 at 20–21 ("The but-for test of causation has therefore not been applied in a rigid way in the common law. Neither has it been considered the exclusive test in negligence cases and courts have recognised that the but-for test needs to be 'supplemented by considerations of justice and legal policy.' Concerned to resolve the unfairness but-for causation can lead to, courts and scholars have admitted that common sense may in some cases require an alternative test of factual causation").

typically functions to limit the civil responsibility of defendants where the injury they cause is too remote or unforeseeable to morally justify recovery.[197] Assessing legal causation, then, is an exercise in policy.[198] In *Mustapha v. Culligan*, the Supreme Court describes a family of cases in which injury factually caused by a defendant's fault is "too remote" to justify recovery.[199] An injury is generally thought to be too remote to be identified as having been legally caused by a defendant's fault if it could not have reasonably been foreseen by a typical person in the defendant's position.[200] Even foreseeable outcomes might sometimes be too remote. Injuries that a reasonable person would think are too "far-fetched" to worry seriously about will not usually be found to have been legally caused by a defendant.[201] Courts generally take the policy view that it would generally be unfair to hold even negligent actors responsible for unimaginable, unpredictable, or uniquely random injuries. Only injuries that normally and foreseeably occur are likely to be legally caused.

b. Civil law causation

Quebec's civil law tradition does not make an explicit distinction between causation in fact and causation in law.[202] Courts are typically satisfied that a defendant's fault caused the plaintiff's injury if the latter was the "certain, direct, and immediate consequence" of the former.[203] In the vast majority

[197] See Khoury, *supra* note 205 at 17 ("This second step aims to determine whether the defendant's act or omission was a sufficiently legally effective cause amongst the complex of other causes. It asks whether the defendant ought to be held liable for the loss suffered by the plaintiff").

[198] David Ozonoff, "Legal Causation and Responsibility for Causing Harm" (2005) 95:1 American Journal of Public Health 35 at 35 ("The other part of legal cause pertains to 'scope of responsibility,' which is usually called 'proximate cause' Proximate cause involves policy and moral questions about the reach of a defendant's liability").

[199] *Mustapha v Culligan of Canada Ltd*, 2008 SCC 27 at para 11, 293 DLR (4th) 29 ("The evidence before the trial judge establishes that the defendant's breach of its duty of care in fact caused Mr. Mustapha's psychiatric injury. We are not asked to revisit this conclusion. The remaining question is whether that breach also caused the plaintiff's damage in law or whether it is too remote to warrant recovery").

[200] Thomasen, *supra* note 148 at 118.

[201] See *Overseas Tankship Ltd (UK) v Miller Steamship Co Pty*, [1967] AC 617 (PC) at 643, [1967] 2 All ER 709.

[202] Khoury, *supra* note 205 at 26 ("No strict divide between factual and legal causation exists").

[203] Khoury, *ibid* at 26; See also, art 1607 CCQ ("The creditor is entitled to damages for bodily, moral or material injury which is an immediate and direct consequence of the debtor's default").

of cases, civil law courts have little difficulty determining whether an injury was caused by a defendant's fault.[204] To the degree there is confusion, it will often stem, as it does in the common law, from fact patterns in which multiple factors combine to cause an injury.[205] Quebec's civil law courts have generally adopted the theory of adequate causation—*causalité adéquate*—in assessing cases of an uncertain causal relationship between fault and injury: adequate causation attempts to isolate the immediate cause of an event, to eliminate mere circumstance in the occurrence of an injury and identify whichever causes would have produced the injury "in a normal state of affairs." [206] An adequate cause, importantly, is often but not always a *sine qua non*, necessary condition for the happening of an event.[207] Adequate causes are usually whichever events, "*dans le cours ordinaire des choses*," would appreciably increase the possibility of an injury occurring.[208] In much the same way that common law courts are unreceptive to the assignment of causation for the highly unlikely, unpredictable consequences of an action, the theory of adequate causation would generally not produce liability for events that are the unusual result of an action. This perspective has important implications for medical malpractice in the use of unexplainable AI. I explain why below.

c. Unexplainable causation

Just as unexplainable AI might complicate the assessment of fault in medical malpractice, so too might it significantly interfere with our usual notions about causation. In this part, I have sketched

[204] Baudouin, Deslauriers & Moore (volume 1), *supra* note 147 at 1-667 ("Dans la plupart des cas, les tribunaux ne soulèvent pas le problème du lien de causalité, parce que la relation entre la faute et le préjudice est évidente").

[205] Baudouin, Deslauriers & Moore (volume 1), *ibid* at 1-667.

[206] Khoury, *supra* note 205 at 27.

[207] Baudouin, Deslauriers & Moore (volume 1), *supra* note 147 at 1-687 ("La jurisprudence québécoise emprunte au système de la causalité adéquate la démarche consistant à séparer la cause véritable des simples circonstances ou occasions du dommage. Ce ne sont donc pas toutes les conditions *sine qua non* qui peuvent et doivent être retenues, mais seulement celles qui ont rendu objectivement possible la réalisation du préjudice.").

[208] Baudouin, Deslauriers & Moore (volume 1), *ibid* at 1-672 ("Pour d'autres, le critère est celui de l'expérience usuelle: est donc une cause adéquate le fait qui, dans le cours ordinaire des choses, accroît sensiblement la possibilité de réalisation du dommage").

out how causation works in common and civil law traditions. Both systems effectively posit that plaintiffs may recover only for injuries factually caused by a defendant and only if the injury is in relatively proximate relation to the fault alleged to have produced it. Reasonably unforeseeable outcomes do not generally admit of recovery. For plaintiffs alleging medical malpractice, causation is often the most challenging element of their claim to prove.[209] Medical facts are frequently uncertain. Evidence may conflict, expert opinions may differ, and confounding factors may be numerous and impossible to conceptually isolate. Unexplainable AI might make each of these challenges worse. There are two broad reasons this might be so. First, unexplainable models meaningfully frustrate our capacity to explain factually how a certain course of events came about. Second, unexplainable models, even if they permit a reliable inference of factual cause, are likely to obscure the proximity of an injury to the fault that produced it.

As we saw in the first chapter above, many of the AI applications most useful in medicine are unexplainable. For those that are *deeply* unexplainable—for which the degree of algorithmic complexity supporting the system precludes even the model's initial programmer from understanding concretely how it works—it may be effectively impossible to give an adequate factual account of the cause of a medical injury. In the imagined scenario that began this part, Dr. *x*'s use of an unexplainable model appears to have caused an injury, Patient *z*'s death. Assuming that Dr. *x*'s fault can be proven or is admitted, the unexplainable character of the sequence of events producing Patient *z*'s injury might yet frustrate recovery for the straightforward reason that some significant portion of the causal story is just not determinable. Courts face uncertain causation in malpractice cases all of the time, but it may be that the functional uncertainty of unexplainable AI poses a uniquely difficult challenge. In some number of cases, an injured plaintiff's evidence will not firmly establish factual causation on a balance of probabilities. This might be so if competing explanations are equally persuasive, such as

[209] See Khoury, *supra* note 205 at 15.

where an injury could just as easily have been brought about by natural causes as by clinical malpractice. This is what happened in the *Snell v. Farrell* case, in which an ophthalmologist operated for the purpose of removing a patient's cataract.[210] During the procedure, the clinician noticed a "a small retrobulbar bleed" but opted to continue with the operation. Several months later, the patient was found to have suffered optic nerve damage and blindness. At trial, the ophthalmologist's decision to proceed with surgery after identifying the patient's bleeding was found to have been fault. But expert testimony could not establish definitively whether the patient's optic nerve damage had been caused by the ophthalmologist's negligence or by natural causes.[211] Justice Sopinka, writing for a unanimous Supreme Court, set out a rule according to which uncertain causation can sometimes be inferred even in the absence of definitive evidence: "The legal or ultimate burden remains with the plaintiff, but in the absence of contrary evidence adduced by the defendant, an inference of causation may be drawn notwithstanding that no positive or scientific proof of causation his available."[212] In *Benhaim v. St. Germain*, the Supreme Court further clarifies that inferences of causation are left to the trier of fact.[213] *Benhaim* affirms *Snell* in requiring that trial courts "take a 'robust and pragmatic' approach to the facts," and allowing "inferences of causation on the basis of 'common sense.'"[214]

---

[210] *Snell*, *supra* note 122 at 315.

[211] *Snell*, *ibid* at 317 ("The plaintiff's expert, Dr. Samis, examined Mrs. Snell in 1985 (about 17 months after the operation) finding new blood vessel formation in the iris, which indicated that she had suffered a stroke in the back of the eye at some point. He could not identify what caused the stroke. He testified that a major cause of optic nerve atrophy is a stroke in the eye itself, which is most likely in a patient with cardiovascular disease, high blood pressure or diabetes. Mrs. Snell suffered from the latter two conditions, although only to the extent that they were controlled by diet rather than medication. Mrs. Snell also suffered from severe glaucoma, which over a long period can also cause optic nerve atrophy. The plaintiff's expert testified that it was unusual to have chronic glaucoma in just one eye, like Mrs. Snell, unless there has been an intervention of some type. The only intervention of which the expert was aware was the operation itself. Neither expert was able to express with certainty an opinion as to what caused the atrophy in this case or when it occurred").

[212] *Snell*, *ibid* at 328–329.

[213] *Benhaim v St. Germain*, 2016 SCC 48 at para 54, [2016] 2 SCR 352 ("The trier of fact may draw an inference of causation even without "positive or scientific proof," if the defendant does not lead sufficient evidence to the contrary. If the defendant does adduce evidence to the contrary, then, in weighing that evidence, the trier of fact may take into account the relative ability of each party to produce evidence").

[214] *Benhaim*, *ibid* at para 54.

Justice Wagner's opinion additionally specifies that even where a defendant's fault is itself a source of causal uncertainty, courts are not obliged to adversely infer causation against the defendant.[215] *Benhaim* notes that the principle set out in *Snell* is an application of rules of evidence set out in article 2849 of the *Civil Code* and in the *Code of Civil Procedure*.

Inferring causation for injury allegedly brought about by the operation of unexplainable AI might be challenging to assess on this robust and pragmatic approach suggested in *Snell* and *Benhaim*. It is not so difficult to imagine conflicting expert evidence, for example, in something like the Patient *z* case above. While a plaintiff expert might reasonably attribute Patient *z*'s death to Dr. *x*'s admitted fault, a defendant expert could compellingly give the opinion that the injury is better attributed to a malfunctioning model, in which case Patient *z*'s injury would have occurred in any event. On the *Snell* and *Benhaim* rules, it is surely open to a trier of fact to draw a causal inference in this case. Both of these evidentiary theories are plausible, and it is well in the function of a trial court to decide which is the more credible.[216] But it is arguable that the unexplainable character of the AI model used in Patient *z*'s care meaningfully muddies the water, making it substantially more difficult on a prudent, common sense approach to understand what precisely caused the plaintiff's injury.[217] It is not, after all, within the ordinary experience of most of us to engage with deeply unexplainable AI, systems

---

[215] *Benhaim*, *ibid* at para 42 ("This Court held in Snell that, in such circumstances, an adverse inference of causation may discharge the plaintiff's burden of proving causation. Those circumstances do not trigger such an inference").

[216] See *British Columbia v Canadian Forest Products Ltd*, 2018 BCCA 124 at para 139, CA43841 ("I conclude that the trial judge did not err in his application of the principles governing causation. After weighing the evidence before him, he concluded that he could not infer that the defendants' negligence probably caused the Province's loss. His finding on this point is entitled to deference. I would not accede to this ground of appeal").

[217] One way of thinking about this problem might be to consider whether malpractice or products liability would be the better venue for Patient *z*'s case. My argument here would suggest that it might not be so easy to know where the case is more coherently brought. Notably, products liability for injury resulting from unexplainable models presents its own problems, for questions about whether and under what conditions the developers of unexplainable models ought to be held liable for the unpredictable effects of their products is not yet a settled issue. See Xavier Frank, "Is Watson for Oncology *per se* Unreasonably Dangerous? Making A Case for How to Prove Products Liability Based on a Flawed Artificial Intelligence Design" (2019) 45:23 American Journal of Law and Medicine 273; Tom Mackie, "Proving Liability for Highly and Fully Automated Vehicle Accidents in Australia" (2018) 34:6 Computer Law & Security Review 1314; Greg Swanson, "Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens" (2019) 42 Seattle University Law Review 1201 at 1204.

that function in ways mysterious even to their creators. Whether a trier of fact is well positioned to infer causation in these conditions is, I think, uncertain. It may be that even expert evidence will be of little assistance, for deeply unexplainable AI is just as opaque to experts as to the court. To be sure, triers of fact will in the face of this uncertainty nevertheless make decisions. Courts will find for plaintiffs where they are convinced a defendant's fault is the likely source of their injury. My point is not that courts will be encumbered in this function. Rather, I want to suggest that there is something logically incoherent, or at least logically dissonant, about this approach. Courts would in awarding damages in these kinds of cases be relying on intuitions to make determinations of fact for which the evidence is not just difficult to understand, but for which we could reasonably say that the evidence is *impossible* to understand. This strikes me as cognitively more difficult to rationalize.

Setting this aside, it may be that unexplainable medical AI also interferes with the law's usual way of thinking about legal causation, particularly in the way that the proximity of a defendant's fault to a plaintiff's injury is evaluated. It might be unclear in a meaningful number of cases whether a clinician using an unexplainable model in patient care could have reasonably foreseen the effects of their relying on the model. This is so to the extent that unexplainable models will sometimes generate surprising, and from the perspective of a human reviewer, unpredictable outcomes.[218] In the balance of cases, unpredictable AI-generated outcomes will improve rather than harm patient care: models might, for example, identify illness or abnormality that a human clinician would have missed. But in some cases, the effect of an unpredicted outcome will be unclear. It may be that precisely in those instances in which a patient is injured and the proximity of the relationship extending from a clinician's conduct to the eventual injury is unclear. To the degree that a model is deeply

[218] See Bathaee, *supra* note 14 at 924 ("In the case of black-box AI, the result of the AI's decision or conduct may not have been in any way foreseeable by the AI's creator or user. For example, the AI may reach a counter-intuitive solution, find an obscure pattern hidden deep in petabytes of data, engage in conduct in which a human being could not have engaged (e.g., at faster speeds), or make decisions based on higher-dimensional relationships between variables that no human can visualize").

unexplainable, it may be that none of its possible outcomes are reasonably foreseeable, at least not in any strict sense. Stated in its strongest terms, I think it is unclear how anyone could reasonably foresee the output effects of a process no one fully understands. One scholar puts it this way: "if even the creator of the AI cannot foresee its effects, a reasonable person cannot either."[219] That a model has produced output results within a certain range in past uses, moreover, does not necessarily ensure that its future output results will fall within that range as well. It might then appear arbitrary for a court to decide that a clinician using some unexplainable model could reasonably have foreseen that the model would have behaved in a particular way or produced a particular output. To hold a clinician, even a negligent or imprudent clinician, responsible for the unpredicted and unpredict*able* consequences of using an otherwise permitted medical implement, might appear patently unjust.

Conclusion

This second chapter outlined how the adoption of unexplainable medical AI might affect malpractice law and, in consequence, the practice and regulation of medicine. I suggested here that unexplainable AI might conceptually interfere the law's prevailing approach to clinical fault and causation. In the first part, I detailed how clinician obligations are determined in Canada's common law and in the civil law of Quebec. I suggested that both are foundationally oriented around the construct of the reasonable professional and that this standard might be difficult to assess when clinicians rely in their practice on unexplainable AI. In the second part, I outlined how courts assess legal and factual causation. I suggested that the straightforward, common-sense approach advocated by Canadian courts is in awkward tension with unexplainable AI. It is hard to know whether a clinician's fault can reasonably be understood to have caused a plaintiff's injury when an intermediating medical model cannot be explained. Of course, none of this should be read to imply that the problems for malpractice

[219] Bathaee, *ibid* at 924.

I describe in this chapter are unresolvable. I gestured at several likely solutions above. Courts and professional bodies will reason through these challenges as they have done with medical innovations of the past. But unexplainable AI nevertheless appears to pose a set of challenges that carry a uniquely conceptually troubling patina. It is no wonder then, that a primary response of jurists and regulators to challenges of this kind has been to suggest that the unexplainable character of our best AI models is a danger the law ought to address. This is the essential supposition underpinning the 'right to explanation' that has in recent years become a cornerstone principle in the regulation of AI. Persons subject to automated decision-making, the argument goes, ought to have a right to understand how decisions affecting their interests were made. In light of the problems raised in this chapter, a right of this kind might have a compelling intuitive appeal. But I argue in the third chapter that a right to an explanation is unlikely to address the problems outlined here and will probably raise significant problems of its own.

CHAPTER THREE: REGULATING EXPLANATION

This third chapter considers whether a right to explanation might remedy the conceptual challenges generated by unexplainable medical AI. I argue that regulating explanation will probably not be effective. It might even make things worse. Rights to explanation have in recent years become a predominant part of formal AI regulation, notably in the European Union's *General Data Protection Regulation* and in Quebec's newly adopted privacy law reform.[220] Though subject to significant critique,[221] rights to explanation are still being enacted. Canada's federal Parliament, for example, is widely expected to consider new privacy legislation sometime in early 2022.[222] That legislation will likely include provisions to enact a national right to explanation for automated decision-making.[223] Rights to explanation might be straightforwardly appealing. Considering some of the conceptual challenges detailed in the second chapter above, it is comprehensible that jurists and policymakers should reflexively favour regulation that would require explainable decision-making. But rights to

[220] See EC, *General Data Protection Regulation* (EU) 2016/679, art 22 & Recital 71, OJ L 119 ("In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision;" "In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision") [GDPR]; Bill 64, *supra* note 16, s 102 ("Any person carrying on an enterprise who uses personal information to render a decision based exclusively on an automated processing of such information must, at the time of or before the decision, inform the person concerned accordingly. He must also inform the person concerned, at the latter's request… of the reasons and the principal factors and parameters that led to the decision").

[221] See Lilian Edwards & Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18 at 22–23 ("After thus taking legal and technological stock, we conclude that there is some danger of research and legislative efforts being devoted to creating rights to a form of transparency that may not be feasible, and may not match user needs. As the history of industries like finance and credit shows, rights to transparency do not necessarily secure substantive justice or effective remedies. We are in danger of creating a 'meaningless transparency' paradigm to match the already well known meaningless consent' trope").

[222] See Murad Hemmadi, "Champagne promises updated privacy legislation in new year" *Regina Leader-Post* (December 6, 2021) ("Innovation Minister François-Philippe Champagne says updating the country's decades-old consumer privacy rules is a 'top priority' in the new Parliament, and he will present 'an amended bill' in the new year after criticism of the Liberal government's original legislation. 'As we're looking at the new economy…this is one way to put Canada at the forefront,' he said in an interview with The Logic on Friday").

[223] Canada's current government introduced privacy reform legislation in in the 43rd Parliament, Bill C-11. That bill prominently included a right to explanation. It died on the Order Paper with the 43rd Parliament's dissolution in August 2020. It appears likely that any attempt to re-introduce a privacy bill in the new Parliament will include a similar right. See Bill C-11, *An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts*, 2nd Sess, 43rd Parl, 2020, s 63(3).

explanation, I suggest in this chapter, miss the mark. I defend this view in three parts. First, I describe how the right to explanation is usually conceptualized by scholars and enacted by legislatures. I define what it purports to guarantee to persons subject to automated decision-making. Second, I suggest how the right to explanation might respond to the challenges for medical practice posed by unexplainable AI. Third, I argue that a right to explanation will likely have a counterproductive impact on medicine.

1. *What the Right to Explanation Guarantees*

No one is accountable for an unexplained decision. This appears to be the intuition that primarily motivates the right to explanation. It also underpins much of how we think about the rule of law in the liberal legal order. An accountable decision-maker explains themselves.[224] This notion of accountability explains why public office holders in (functioning) democratic systems of government are generally expected by the electorate to explain and justify their policy choices.[225] Judges give written opinions when deciding cases[226] and administrative officials are bound by procedural fairness to provide written reasons for many of the decisions they make.[227] Accountability on these terms

[224] See Glen Staszewski, "Reason-Giving and Accountability" (2009) 93 Minnesota Law Review 1253 at 1279 ("The word 'accountable' means that one is 'required or expected to justify actions or decisions.' Although dictionary definitions, standing alone, should certainly not always be dispositive, my contention is that public officials in a democracy can be held deliberatively accountable by a requirement or expectation that they give reasoned explanations for their decisions").

[225] Mark Philp, "Delimiting Democratic Accountability" (2009) 57 Political Studies 28 at 32 ("Modern democracies rest on a combination of two ideas: that those who rule should do so in the public interest or in response to the public will; and that they will be more likely to do so when they are, in some way, representative of, and/or accountable to those they rule. This much is uncontentious among most democrats").

[226] Claire L'Heureux-Dubé, "The Length and Plurality of Supreme Court of Canada Decisions" (1990) 28 Alberta Law Review 581 at 585 ("Why are written opinions occasionally so lengthy and numerous? The answer lies in the role which the judges set for themselves. It is easy to sit back and vote; it is another matter to write persuasive reasons that will gather support for the conclusion; and it is yet more difficult to render justice while at the same time helping to shape the law of tomorrow"); See also Benjamin Eidelson, "Reasoned Explanation and Political Accountability in the Roberts Court" (2021) 130 Yale Law Journal 1748.

[227] Grant Huscraft, "From Natural Justice to Fairness: Thresholds, Content, and the Role of Judicial Review" in Colleen M Flood & Lorne Sossin, eds, *Administrative Law in Context*, 2d ed (Toronto: Emond, 2013) 147–184 at 177; *Baker v Canada (Minister of Citizenship and Immigration)*, [1999] 2 SCR 817, 174 DLR (4th) 193 at para 43 ("In my opinion, it is now appropriate to recognize that, in certain circumstances, the duty of procedural fairness will require the provision of a written explanation for a decision").

consists in 'accounting for' or in 'giving an account' of certain kinds of conduct: "A is accountable with respect to M when some individual, body or institution, Y, can require A to inform and explain/justify his or her conduct with respect to M."[228] Of course, we generally do not refer only to explanation and justification when we talk about accountability. Being 'held to account' might in some cases have a punitive or restorative connotation: we might reasonably say, for example, that someone convicted of an offense is made accountable in the discharge of a criminal sentence. But explanation is at least a *part* of what we mean. Unexplained decisions can seem arbitrary, or worse, unfair. Decisions unaccompanied by justification cannot easily be reviewed or contested. Persons subject to unexplained decisions might reasonably feel they were submitted to an unjust process.

It is effectively this absence of accountability to which the right to explanation for automated decision-making tries to respond. By compelling the developers and users of unexplainable AI models to explain how and why the systems reach the decisions that they do, the right to explanation attempts to respond to growing scholarly demand for AI accountability.[229] In some places, the demand is highly prospective, with academics speculating on legal developments not yet adopted by bureaucrats or legislatures.[230] In others, most notably in the European Union, an entire regulatory infrastructure has been in place for some time. But in either case, explanation is generally thought to be a critical part of policy frameworks advancing AI accountability.[231] Some authors note that this focus of jurists and

[228] Philp, *supra* note 247 at 32 ("The definition has four dimensions: (1) the agent or institution who is to give an account; (2) the agent or institution to whom or to which they give an account; (3) the responsibilities or domain of actions that are the subject matter of the account they give; and (4) the capacity of Y to require A to give an account – so that (4) defines the following relationship, that (1) can be required to inform and explain or justify his or her actions regarding (3) to (2)").

[229] See Deven R Desai & Joshua A Kroll, "Trust but Verify: a Guide to Algorithms and the Law" (2017) 31 Harvard Journal of Law & Technology 1 at 7 ("Both legal-political and computer science scholars wish to ensure that automated decision systems are not enabling misdeeds or generating undesired outcomes. Both fields use the terms transparency and accountability, but have different meanings and functions for them. This vocabulary clash can muddy the understanding of what is desired as an end result and of how to create systems to realize whatever end result is agreed upon").

[230] See Margot E Kaminski, "The Right to Explanation, Explained" (2019) 34 Berkeley Technology Law Journal 189 at 190–191 ("The literature in the United States has been largely speculative, operating in a policy vacuum").

[231] See e.g. Ashley Deeks, "Rulemaking and Inscrutable Automated Decision Tools" 119 Columbia Law Review 1851 at 1856–1857 ("Explanation requirements, including a duty to inform principals of facts that 'the principal would wish to have' or 'are material to the agent's duties,' are basic mechanisms for ensuring that agents are accountable to principals…

policymakers on explanation might reveal unrealistic attitudes about the AI models to which these new rules are intended to apply.[232] Others have described unexplainable AI as having induced a kind of moral panic among policymakers. Faced with what was assumed to be an unprecedented legal, ethical, and technological challenge, regulators reflexively proposed explanation as a straightforward solution: "any remedy in a storm has looked attractive."[233] Though several scholars have argued (I think persuasively) that explanation is a poorly constructed way of achieving AI accountability, it continues to be adopted by legislatures and defended by policymakers. I will return in later parts of the chapter to explanation's status as a hastily constructed, ineffective solution to the kinds of problems AI accountability advocates have sought to avoid. In this first part, I examine how rights to explanation have been implemented in two jurisdictions: the European Union and Quebec. I also consider recently proposed amendments to Canada's federal private sector privacy law.

a. European Union: Article 22 and Recital 71 of the GDPR

Though the European Union's *General Data Protection Regulation* (GDPR) is effectively the conceptual ancestor of most of the existing statutory rights to explanation, Quebec's included, whether the regulation even contains a right of this kind was initially a matter of some controversy.[234]

Expanding the focus of the explainability debate to include public accountability is thus only one step toward a more realistic view of the ramifications of decision tool inscrutability").

[232] Desai & Kroll, *supra* note 251 at 4 ("Put simply, current calls for algorithmic transparency misunderstand the nature of computer systems").

[233] See Edwards & Veale, *supra* note 243 at 81 ("Transparency in the form of a 'right to an explanation' has emerged as a compellingly attractive remedy since it intuitively presents as a means to 'open the black box,' hence allowing individual challenge and redress, as well as possibilities to foster accountability of ML systems. In the general furore over algorithmic bias, opacity and unfairness laid out in section I, any remedy in a storm has looked attractive").

[234] See Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation'" (2017) 38:3 AI Magazine 50 at 55 ("Although the article does not elaborate what these safeguards are beyond 'the right to obtain human intervention,' Articles 13 and 14 state that, when profiling takes place, a data subject has the right to "meaningful information about the logic involved.' This requirement prompts the question: what does it mean, and what is required, to explain an algorithm's decision?"); Sandra Wachter, Brent Mittelstadt & Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2017) 7:2 International Data Privacy Law 76 at 77 ("There are several reasons to doubt the existence, scope, and feasibility of a 'right to explanation' of automated decisions. In this article, we examine the legal status of the 'right to explanation' in the GDPR, and identify several barriers undermining its implementation. We argue that the GDPR does not, in its current form, implement a right to explanation, but rather what we term a limited 'right to be informed'").

That the GDPR might include a right to explanation for persons subject to certain kinds of automated decision-making came as a considerable "surprise to some EU data protection lawyers," for no right to explanation is mentioned explicitly or directly identified in the regulation's text.[235] Proponents of the existence of the right to explanation ground their case on a contextual reading of multiple articles, a particular interpretation of the GDPR's legislative history, and on at least one of the regulation's Recitals—a non-binding explanatory legislative note.[236] Nearly everyone agrees that Article 22 of the GDPR is essential for understanding whether a right to explanation exists and, if it does, for understanding what it guarantees. The relevant portions of Article 22 are the following:

> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
> 2. Paragraph 1 shall not apply if the decision:
>   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>   (c) is based on the data subject's explicit consent.
> 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.[237]

On its face, Article 22 creates a default prohibition on automated decision-making. It at least appears to create a right to object on the part of data subjects about whom automated decisions are made.[238]

---

[235] Edwards & Veale, *supra* note 243 at 44 ("In 2016, to the surprise of some EU data protection lawyers, and to considerable global attention, Goodman and Flaxman asserted in a short paper that the GDPR contained a 'right to an explanation' of algorithmic decision making. As Wachter et al. have comprehensively pointed out, the truth is not quite that simple").

[236] Kaminski, *supra* note 252 at 196 ("Article 22 of the GDPR addresses '[a]utomated individual decision-making, including profiling.' Articles 13, 14, and 15 each contain transparency rights around automated decision-making and profiling. More general GDPR provisions, such as the right to object, the right to rectification (correction), data protection by design and by default, and the requirement of data protection impact assessments, likely apply to most or even all algorithmic decision-making").

[237] GDPR, *supra* note 242, art 22.

[238] See Kaminski, *supra* note 252 at 196 ("Scholars have pointed out, based on the historical treatment of similar text in the Data Protection Directive (DPD), the predecessor to the GDPR, that this could be interpreted as either a right to object to such decisions or a general prohibition on significant algorithmic decision-making").

In its first paragraph, the article limits its scope of application in two important respects. First, only automated decisions based *solely* on automated processes are banned. That a decision is based solely on automated processing is generally understood to mean that no person "exercises any real influence on the outcome of the decision-making process."[239] This would suggest that decisions produced by an AI model in support of human decision-making likely would fall beyond Article 22's reach. In the view of the European Data Protection Board, the consensus body tasked with interpreting and ensuring uniform application of the GDPR across the European Union, a minimal level of human involvement, such as cursory review or rubber-stamping, would be insufficient to omit a decision-making process from oversight under this section of the GDPR.[240] It is likewise probable that decisions otherwise based solely on automated processing, but which have been merely attributed to a human absent input or review would not escape the scope of this article.[241]

Second, the prohibition in Article 22 appears to apply only to automated decisions that "produce legal effects" or are "similarly significant" in their impact.[242] Scholars seem to agree that a decision having legal effects on an individual would at minimum include decisions affecting their

[239] Isak Mendoza & Lee A Bygrave, "The Right Not to be Subject to Automated Decisions Based on Profiling" in Tatiana-Eleni Synodinou et al, eds, *EU Internet Law: Regulation and Enforcement* (Cham: Springer, 2017) 77–97 at 87 ("This leads to the second condition which is that the decision is based solely on automated data processing. By this is meant that a person fails to exercise any real influence on the outcome of the decision-making process. Even if a decision is formally ascribed to a person, it is to be regarded as based solely on automated processing if a person does not actively assess the result of the processing prior to its formalisation as a decision").

[240] Article 29 Data Protection Working Party, "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" (EU) 2017/WP251 ("The controller cannot avoid the Article 22 provisions by fabricating human involvement. For example, if someone routinely applies automatically generated profiles to individua ls without any actual influence on the result, this would still be a decision based solely on automated processing. To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data") [European Data Protection Board]; Kaminski, *supra* note 252 at 197 ("One could narrowly interpret "based solely" to mean that any human involvement, even rubber-stamping, takes an algorithmic decision out of Article 22's scope; or one could take a broader reading to cover all algorithmically-based decisions that occur without meaningful human involvement").

[241] Mendoza & Bygrave, *supra* note 261 at 87 ("Even if a decision is formally ascribed to a person, it is to be regarded as based solely on automated processing if a person does not actively assess the result of the processing prior to its formalisation as a decision").

[242] GDPR, *supra* note 242, art 22(1).

legal status, for example on existing legal entitlements in civil procedure or administrative law.[243] On this interpretation, Article 22's prohibition on automated decision-making might extend to decisions that would tend to modify a data subject's constitutional entitlements, such as a right to health, freedom of expression, or freedom of movement.[244] But it is unclear on a first reading whether the legal relationship between a data custodian and data subject matters in our appreciation of whether automated processing produces legal effects. Whether a private entity's use of automated processing in a way that affects the legal entitlements of an individual owed by the state falls within this article's ambit, for example, is not obviously settled in the text of the GDPR. It is also unclear precisely what kinds of decisions this language of legal effect is meant to exclude. While certain automated decisions might have a significant impact on a data subject's interests, such as decisions about employment or targeted advertising, Article 22 does not itself resolve whether impact on these interests are sufficiently like legal effects to be regulated under this provision.[245] On the European Data Protection Board's interpretation, data processing significantly affects an individual if "the effects of the processing must be sufficiently great or important to be worthy of attention."[246] While the Board

[243] Gianclaudio Malgieri & Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation" (2017) 7:4 International Data Privacy Law 243 at 252 ("In particular, it has been argued that having legal effects on data subjects means 'affecting their legal status.' In other words, any influence (eg limitation) on existing rights of data subjects, including civil procedural rights, administrative law rights, etc is a legal effect. Analogously, we can infer that any influence on human rights or constitutional rights of individuals is a legal effect").

[244] Malgieri & Comandé, *ibid* at 252 ("Analogously, we can infer that any influence on human rights or constitutional rights of individuals is a legal effect. Thus, any limitation to, eg, the right to health, freedom of expression, freedom of movement, right not to be illegitimately discriminated, etc is a legal effect. Incidentally, we recall that human rights enlisted in the European Convention of Human Rights and in the EU Charter of Fundamental Rights cannot be considered a 'numerus clausus'").

[245] Malgieri & Comandé, *ibid* at 252 ("An example can be recruitment: there is not any right to be accepted for a job position, but there is a legitimate interest to be assessed loyally, fairly and not to be dis- criminated during recruitment. Another example is online behavioural advertisements exploiting consumers' biases: there is not a right not to receive personalized advertisements, but there is a legitimate interest not to be the victim of unfair commercial practices or cognitive manipulation").

[246] European Data Protection Board, *supra* note 230 at 21 ("For data processing to significantly affect someone the effects of the processing must be sufficiently great or important to be worthy of attention. In other words, the decision must have the potential to: significantly affect their circumstances, behaviour or choices of the individuals concerned; have a prolonged or permanent impact on the data subject; or at its most extreme, lead to the exclusion or discrimination of individuals").

acknowledges the difficulty in compiling a complete and exhaustive list of factors that would be triggered by this provision, it notes generally that decisions likely to have an impact on a person's financial circumstances, access to health services, or access to employment, likely would be captured by the 'similarly significant effects' clause.[247]

In the article's second paragraph, three further limitations on the application of the prohibition on automated decision-making are described. Decision-making based on an automated process is permissible, for example, where it is necessary for the execution of a contract, is authorized by adequately protective state law, or is based on a data subject's explicit consent.[248] These are conditions reflected elsewhere in the GDPR and which extend throughout the regulation's legal architecture.[249] Even where one or more of these exceptions apply and automated decision-making is consequently not prohibited by Article 22(1), the article's third paragraph requires that data controllers take "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests."[250] This clause minimally requires that a data subject have the option to "obtain human intervention," to express their point of view, or contest the decision in question.[251] This 'suitable measures' clause is, for scholars who think the GDPR expresses a right to explanation, where the right is grounded.[252] Of

[247] European Data Protection Board, *supra* note 230 at 21 ("It is difficult to be precise about what would be considered sufficiently significant to meet the threshold, although the following decisions could fall into this category: decisions that affect someone's financial circumstances, such as their eligibility to credit; decisions that affect someone's access to health services; decisions that deny someone an employment opportunity or put them at a serious disadvantage").

[248] GDPR, *supra* note 242, art 22(2).

[249] See e.g. GDPR, *ibid*, art 9 ("Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited…Paragraph 1 shall not apply if one of the following applies: the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject").

[250] GDPR, *supra* note 242, art 22(3).

[251] GDPR, *ibid*, art 22(3).

[252] Kaminski, *supra* note 252 at 198 ("Even when an exception to Article 22 applies, a company must implement 'suitable measures to safeguard the data subject's rights and freedoms and legitimate interests…' This requirement is the source of the debate over the right to explanation").

course, the right to "an opportunity to be heard," which is all that Article 22(3) appears to explicitly require,[253] is a fairly shallow conception of what we might call algorithmic due process. But the specific measures identified in the article might not be all that the suitable measures passage could conceivably demand. In fact, that the passage itself guarantees "at least" the right to an opportunity to be heard seems to indicate that more is expected of data controllers than this explicitly contemplated minimum level of protection.[254]

This perspective is in principle supported both by the European Data Protection Board's interpretive comments on Article 22 and by Recital 71 to the GDPR. These documents suggest, and most scholars agree, that though Article 22 anchors the right to explanation, it must be read alongside Articles 13–15. In Articles 13 and 14, the GDPR creates a sweeping right for data subjects to be notified whenever information about them is collected.[255] It also creates, in Article 15, a right to be informed of "the existence of automated decision-making," and to be provided "meaningful information about the logic involved."[256] These articles have been the subject of significant debate, particularly about the nature and timing of the information they each require to be communicated to

[253] Kaminski, *supra* note 252 at 198 ("This explicitly creates a version of algorithmic due process: a right to an opportunity to be heard. These are the only safeguards named in the GDPR's text").

[254] GDPR, *supra* note 242, art 22(3); Kaminski, *supra* note 252 at 198 ("The use of the words "at least," however, indicates that these are an open list of minimum requirements, and a company should do more. As discussed in Part IV, both the preamble (Recital) and interpretative guidance have added to this list of both suggested and required safeguards, and both include as a safeguard a right to explanation of an individual decision").

[255] See GDPR, *ibid*, arts 13(2)(f) & 14(2)(g); European Data Protection Board, *supra* note 230 at 24–25 ("Given the potential risks and interference that profiling caught by Article 22 poses to the rights of data subjects, data controllers should be particularly mindful of their transparency obligations. Articles 13(2) (f) and 14(2) (g) require controllers to provide specific, easily accessible information about automated decision-making, based solely on automated processing, including profiling, that produces legal or similarly significant effects").

[256] See GDPR, *ibid*, art 15(1)(h) ("The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject").

data subjects.[257] Some authors stress that the 'meaningful information' clause in Article 15 appears to require little more than the *ex-ante* conveyance of very basic information about the way a model used in automated decision-making functions.[258] But others suggest a more expansive interpretation. When read in the GDPR's broader context, and particularly in conjunction with Article 22, the 'meaningful interpretation' clause seems to admit of multiple meanings.[259] This perspective is echoed in the GDPR's accompanying explanatory texts. In its guidelines on automated decision-making, for example the European Data Protection Board suggests that the GDPR requires that data controllers "provide meaningful information about the logic involved" in reaching an automated decision, but not necessarily "a complex explanation of the algorithms used or disclosure of the full algorithm."[260] At the same time, the Board specifies that information communicated under Article 15 ought to be "sufficiently comprehensive for the data subject to understand the reasons for the decision."[261] Recital 71 is more direct. Informed by Article 15's meaningful interpretation clause, it appears to find a right to explanation in the suitable measures clause of Article 22(3):

> In any case, [automated] processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.[262]

---

[257] See Kaminski, *supra* note 252 at 199 ("This language has provoked debate, especially over the question of timing. The language in all three Articles is identical, but the temporal context is different. Articles 13 and 14, roughly speaking, require companies to notify individuals when data is obtained, while Article 15 creates access rights at almost any time").

[258] Wachter, Mittelstadt & Floridi, *supra* note 256 at 90 ("Paal also notes that the purpose of Article 15 GDPR is to allow data subjects to be informed about the usage and functionality of auto- mated decision-making. As the scope of information data controllers are required to disclose in Article 15 is the same as in Article 13, Article 15 similarly requires only limited information about the functionality of the automated decision-making system").

[259] Malgieri & Comandé, *supra* note 233 at 245 ("In particular, we will explain that legibility offers the most appropriate interpretation of the right to know 'meaningful information about the logic involved in a decision-making' (Article 15(1)(h) combined with Article 22 of GDPR)").

[260] European Data Protection Board, *supra* note 230 at 25.

[261] European Data Protection Board, *ibid* at 25.

[262] GDPR, *supra* note 242, Recital 71.

Though this is as explicit as the GDPR gets in generating a right to explanation, it is important to note, as I mentioned above, that Recital 71 is not by itself legally binding—only an article can create a properly legally enforceable right.[263] It might appear mysterious that the right predicted in Recital 71 would exist in Article 22 without being mentioned directly in the latter. If a right to explanation indeed exists in Article 22, then it is secondary to the right not to be subject to automated decision-making except under specific conditions. Recital 71's right to explanation would apply, in other words, only to automated decision-making made in one of the exceptions described in Article 22(2), such as with the explicit consent of the data subject.[264] All of this leaves the right to explanation in a curious position in the GDPR's broader regulatory framework. Because it is mentioned explicitly only once in the entire text of the GDPR, in a technically non-binding preambulatory note, it is difficult to know exactly how binding the right is supposed to be.[265] Some authors suggest that this uncertain way of addressing the right to explanation is primarily the result of political expediency: "issues too controversial for agreement in the main text have been kicked into the long grass of the recitals."[266] Setting aside the interpretive challenges this causes, this is an approach that characterizes much of the GDPR. This is, in the view of many scholars, an inherently and purposively collaborative

[263] See Mendoza & Bygrave, *supra* note 261 at 92–92 ("Express mention of this right in the Regulation occurs solely in Recital 71, which lists the right in its elaboration of 'suitable safeguards.' However, a Recital does not of itself create a legally binding right; the latter may only be created pursuant to an Article").

[264] Edwards & Veale, *supra* note 243 at 49 ("Recital 71 then mentions all of the above safeguards but also adds a further, explicit "right to an explanation." Is this therefore another route to a "right to an explanation" in Article 22? This seems paradoxical. Article 22 gives a primary right, i.e. to stop wholly automated decision making. Would it give what seems an equally powerful right—to an explanation—in circumstances where the primary right is excluded because the data subject has already consented to the processing?").

[265] See Wachter, Mittelstadt & Floridi, *supra* note 256 at 77–78 ("The aforementioned claim for a right to explanation muddles the first and second legal bases. It conflates (i) legally binding requirements of Article 22 and non-binding provisions of Recital 71 and (ii) notification duties (Articles 13–14) that require data subjects to be provided with information about 'the *existence* of *automated decision-making*, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, *meaningful information* about the *logic involved*, as well as the *significance* and the *envisaged consequences* of such processing for the data subject' (emphasis added)").

[266] Edwards & Veale, *supra* note 243 at 50.

regulation, a document that is supposed to evolve with time.[267] On that view, unclarity about what the right to explanation will mean in practice, and indeed whether a right to explanation as a binding component of the GDPR exists at all, might be an intentional feature of the GDPR's approach to regulating automated decision-making.

What this all means, of course, is that it is ultimately up to the twenty-seven individual EU member states through domestic statutory law and up to the courts to decide how the explanatory obligations of data controllers ought to be governed. EU member states have since the adoption of the GDPR in 2016 enacted a substantively diverse set of national data protection statutes. Most of them have not adopted an explicit right to explanation.[268] Notable exceptions include France and Hungary. French law, for example, provides a right to explanation on the request of an individual subject to an automated decision under Article 22 of the GDPR defined in terms of the rules according to which an individual's data is processed and the principal features of the relevant automated system's operation.[269] At the level of case law interpreting Article 22, European courts have to date been fairly silent.[270] A Dutch case decided in the spring of 2021 decided, to my knowledge for the

[267] Kaminski, *supra* note 252 at 195 ("Thus, when scholars argue that what is in the Recitals is not the law,21 they are not only insisting on a technicality—distinguishing between harder and softer legal instruments—they are also disregarding the fundamentally collaborative, evolving nature of the GDPR, and removing important sources of clarity for companies as the law develops").

[268] Gianclaudio Malgieri, "Automated decision-making in the EU Member States: The right to explanation and other 'suitable safeguards' in the national legislations" (2019) 35 Computer Science Law & Security Review 1 at 22 ("Most Member States do not include [safeguards for automated decision-making] in their national data protection law. The only exceptions are Hungary and France: in both these cases, such right is based on the request of the data subject, but the requirements of this explanation are slightly different").

[269] *Loi no 2018-493 du 20 juin 2018 relative à la protection des données personnelles*, JO, 21 June 2018, JUSC1732261L, art 21(I)(1º) ("Des cas mentionnés aux a et c du 2 de l'article 22 du règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 précité, sous les réserves mentionnées au 3 du même article 22 et à condition que les règles définissant le traitement ainsi que les principales caractéristiques de sa mise en œuvre soient communiquées, à l'exception des secrets protégés par la loi, par le responsable de traitement à l'intéressé s'il en fait la demande").

[270] See Raphaël Gellert, Marvin van Bekkum & Frederik Zuiderveen Borgesius, "The Ola & Uber judgments: for the first time a court recognises a GDPR right to an explanation for algorithmic decision-making" (2021) EU Law Analysis, online: <http://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html> ("In the Ola judgment, the Court requires the Ola company to explain the logic behind a fully automated decision in the sense of article 22 of the General Data Protection regulation (GDPR). This is the first time that a court in the Netherlands recognises such a right. To the best of our knowledge, it is also the first time that a Court anywhere in Europe recognises such a right").

first time in the EU, that the GDPR in fact does contain a right to explanation. In the *Ola Netherlands* case, the Amsterdam District Court ruled that a data controller was required to "communicate the main assessment criteria and their role in the automated decision to [the applicants], so that they can understand the criteria on the basis of which the decisions were taken and they are able to check the correctness and lawfulness of the data processing."[271] Ola, a rideshare service similar to Uber or Lyft, developed a driver monitoring AI model. Ola used the system to monitor driver behaviour, detect service irregularities, and impose fines in cases of suspected fraudulent driver conduct.[272] Decisions to fine drivers would occur automatically, without human direct human oversight. Ola's model produced decisions that, in the reasoning of the District Court, are significant enough to warrant attention under Article 22(1) of the GDPR, especially insofar as the automated process imposes sanctions that affect the (legal) rights of drivers to remuneration under their work agreement with Ola.[273] These factors together trigger the GDPR's general prohibition on automated decision-making, an outcome that not tempered by the exemptions set out in Article 22(2). In particular, the District Court finds that it is impossible to determine "whether Ola has taken appropriate safeguards within the meaning of paragraph 3 of Article 22 GDPR."[274] This decision, decided relatively early in the jurisprudential life of Article 22, does not definitively settle that Article 22 of the GDPR ought to be interpreted to include a right to explanation, much less what the contours of such a right could be. It

[271] Amsterdam District Court, Amsterdam, 11 March 2021, *Applicant v Ola Netherlands BV*, C/13/689705, HA RK 20-258 at para 4.52.

[272] *Ola Netherlands BV*, *ibid* at paras 4.48–4.51 ("Ola uses the Guardian system to detect irregularities. …with regard to Ola's automated decision-making process that determines that rides are not valid and that, as a result, discounts or fines ('penalties and deductions') are imposed").

[273] *Ola Netherlands BV*, *ibid* at para 4.51 ("Contrary to what Ola argues, the decision to impose a discount or fine has effects that are important enough to merit attention and that significantly affect the behavior or choices of the person concerned as referred to in the Guidelines. After all, such a decision leads to a sanction that affects the rights of [applicants] under the agreement with Ola").

[274] *Ola Netherlands BV*, *ibid* at para 4.51 ("As a result, it is not possible to answer the question whether Ola has taken appropriate safeguards within the meaning of paragraph 3 of Article 22 GDPR").

suggests, though, that courts, guided by Recital 71 and the interpretation of the European Data Protection Board, are likely at least open to ordering the purveyors of automated decisions to explain themselves to persons affected by the judgements of AI models.

b. Quebec: Bill 64, An Act to modernize the protection of personal information

In September 2021, Quebec's National Assembly unanimously adopted Bill 64, perhaps the most significant reformation of privacy law undertaken anywhere in Canada in a generation.[275] Introducing the bill on June 12, 2020, then Minister of Justice Sonia LeBel described the function of Bill 64 as "modernizing the framework applicable to the protection of information in a number of statutes" and "introducing laws concerning the handling of incidents affecting the confidentiality of personal information by public bodies and businesses."[276] Bill 64's effects are wide-ranging. It creates new obligations on the part of certain private sector organizations, for example, to appoint privacy officers responsible for overseeing compliance with the *Act*,[277] to report data breaches that pose serious risk of injury to individuals,[278] and to undertake privacy impact assessments for the acquisition, use, or

[275] See Bill 64, *supra* note 16; Rish Handa, "Bill 64 Enacted: Québec's Modern Privacy Regime" *McMillan Business Law and Regulatory Law Bulletin* (October 15, 2021), online: <https://mcmillan.ca/insights/bill-64-enacted-quebecs-modern-privacy-regime/> ("On September 21st, the National Assembly of Québec unanimously voted to pass Bill 64, an *Act to modernize legislative provisions as regards the protection of personal information*, and the bill received assent the following day. The result is that Québec has significantly modernized its private and public sector privacy regimes, better adapting its legislative framework for the protection of personal information to present-day realities and keeping pace with international privacy developments").

[276] Quebec, National Assembly, *Journal des débats (Hansard)*, 42-1, vol 45, no 120 (12 juin 2020) at 8245 ("Ce projet de loi modernise l'encadrement applicable à la protection des renseignements personnels dans diverses lois… Le projet de loi introduit à ces deux lois des règles concernant le traitement des incidents affectant la confidentialité des renseignements personnels par les organismes publics et les entreprises").

[277] Bill 64, *supra* note 16, art 95 ("Within the enterprise, the person exercising the highest authority shall see to ensuring that this Act is implemented and complied with. That person shall exercise the function of person in charge of the protection of personal information; he may delegate all or part of that function in writing to a personnel member").

[278] Bill 64, *ibid*, art 95 ("Any person carrying on an enterprise who has cause to believe that a confidentiality incident involving personal information the person holds has occurred must take reasonable measures to reduce the risk of injury and to prevent new incidents of the same nature. If the incident presents a risk of serious injury, the person carrying on an enterprise must promptly notify the Commission d'accès à l'information established by section 103 of the Act respecting Access to documents held by public bodies and the Protection of personal information").

transfer of specific kinds of personal information.[279] Public entities are similarly subject to a broad array of new obligations, among them a requirement to assess the "privacy-related factors of any information system project or electronic service delivery project involving the collection, use, release, keeping or destruction of personal information."[280] Though modelled plainly on the GDPR,[281] Bill 64 diverges from its European cousin in that it leaves no significant ambiguity about whether it has intentionally created a right to explanation—it arguably creates two of them. These rights are drafted as new obligations on the part of private entities and public bodies that use personal information for making decisions on the basis of automated processing. Bill 64 places these new rights to explanation in two extensively amended statutes: the *Act Respecting the Protection of Personal Information in the Private Sector* and the *Act Respecting Access to Documents Held by Public Bodies and the Protection of Personal Information*. Though applicable in different worlds, the public and private sector rights to explanation adopt substantially the same language, reproduced here:

> A [public body or any person carrying on an enterprise] that uses personal information to render a decision based exclusively on an automated processing of such information must, at the time of or before the decision, inform the person concerned accordingly.
>
> It must also inform the person concerned, at the latter's request,
> (1) of the personal information used to render the decision;
> (2) of the reasons and the principal factors and parameters that led to the decision; and

[279] Bill 64, *ibid*, arts 95 & 103 ("Before communicating personal information outside Québec, a person carrying on an enterprise must conduct an assessment of privacy-related factors. The person must, in particular, take into account (1) the sensitivity of the information; (2) the purposes for which it is to be uses; (3) the protection measures that would apply to it; and (4) the legal framework applicable in the State in which the information would be communicated").

[280] Bill 64, *ibid*, art 14 ("A public body must conduct an assessment of the privacy-related factors of any information system project or electronic service delivery project involving the collection, use, release, keeping or destruction of personal information. For the purposes of such an assessment, the public body must consult its committee on access to information and the protection of personal information from the outset of the project. The public body must also ensure that the project allows computerized personal information collected from the person concerned to be released to him in a structured, commonly used technological format").

[281] See Quebec, National Assembly, *Journal des débats of the Committee on Institutions (Hansard)*, "Clause-by-clause consideration of Bill 64, An Act to modernize legislative provisions as regards the protection of personal information" 42-1, vol 45, no 113 (2 février 2021) at 10h ("On nous dit: Bien voyons donc! C'est un magnifique projet de loi, on se base sur — puis vous me direz, M. le Président, combien de temps qu'il me reste — on se base sur le règlement européen, qui a été adopté et qui donne le ton, qui est un magnifique règlement, très bien fait").

(3) of the right of the person concerned to have the personal information used to render the decision corrected.[282]

This statutory language bears an obvious resemblance to that included in Article 22 of the GDPR. Most noteworthy is that Quebec's right to explanation applies to decision-making based *exclusively* on an automated processing of personal information, just as its European counterpart considers only decisions *solely* derived from automated processes.[283] But Bill 64's provisions also differ meaningfully from what appears in the text of the GDPR. For one thing, Bill 64 takes a clear position on the timing of the access of an individual to an explanation of the decision affecting them. Whereas the GDPR framework apparently takes no precise view on when information surrounding automated decision-making needs to be communicated with data subjects,[284] the right to explanation in Bill 64 extends at a specific juncture: before or at the time a decision is made. This approach differs from that implied by Recital 71 of the GDPR, which suggests that the right to explanation applies only *after* the making of an automated assessment.[285] For another thing, Bill 64 does some work toward specifying what ought to be contained in an explanatory communication. In particular, Quebec notes that individuals subject to a decision should be provided the "reasons and the principal factors and parameters" that led to a decision. In its French text, the *Act* employs a similar construction, with the right to explanation applying to "*des raisons, ainsi que des principaux facteurs et paramètres, ayant mené à la decision*."[286] Interpreting its plain language meaning, this way of approaching the right to

[282] Bill 64, *supra* note 16, arts 20 & 102.

[283] See GDPR, *supra* note 242, art 22 ("The data subject shall have the right not to be subject to a decision based *solely* on automated processing").

[284] See Kaminski, *supra* note 252 at 216 ("This language has provoked debate, especially over the question of timing. The language in all three Articles is identical, but the temporal context is different. Articles 13 and 14, roughly speaking, require companies to notify individuals when data is obtained, while Article 15 creates access rights at almost any time").

[285] GDPR, *supra* note 242, Recital 71 ("In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached *after such assessment* and to challenge the decision" emphasis mine).

[286] *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*, CQLR, c 25.

explanation in Bill 64 implies a demand for three distinct varieties of explanation: reasons, principal factors, and parameters.[287] None of these concepts are defined and it is unclear if any interpretive differences ought to be drawn between them. It is thus not immediately apparent what precisely the right to explanation in Bill 64 is trying to protect. It will be up to the courts to settle what kinds of reasons, factors, or parameters ultimately satisfy this newly created right to explanation.

c. Canada: PIPEDA reform, Bill C–11

Quebec's Bill 64 is the first significant privacy law reform undertaken in Canada since the adoption of the GDPR. Other provinces[288] and the federal government have been to varying degrees moving toward modernizing privacy statutes in the model of European Union. Federal privacy law reform has in recent years been mired in political controversy. Academic and private sector critique, not to mention the objections of the Privacy Commissioner of Canada, have characterized recent efforts to amend Canada's primary private sector privacy statute, the *Personal Information Protection and Electronic Documents Act*, PIPEDA.[289] Navdeep Bains, then Minister of Science, Innovation, and Industry, introduced Bill C–11, the *Digital Charter Implementation Act*, in the House of Commons on November 17, 2020.[290] Bill C–11 immediately attracted controversy. Commenters noted that the

[287] Notably, the French and English texts are equally authoritative. See *Charter of the French Language*, CQLR, c 11, s 7 ("French is the language of the legislature and the courts in Québec, subject to the following: (1) legislative bills shall be printed, published, passed and assented to in French and in English, and the statutes shall be printed and published in both languages…(3) the French and English versions of the texts referred to in paragraphs 1 and 2 are equally authoritative").

[288] See e.g. Ontario, White Paper, *Modernizing Privacy in Ontario Empowering Ontarians and Enabling the Digital Economy* (Toronto: Government of Ontario, 2021).

[289] See Christopher Guly, "Data-protection bill needs to be on parliamentary agenda this session, says privacy-law expert," *the Hill Times* (27 October 2021), online: <https://www.hilltimes.com/2021/10/27/data-protection-bill-needs-to-be-on-parliamentary-agenda-this-session-says-privacy-law-expert/324718> ("Last November, then-innovation, science, and industry minister Navdeep Bains tabled before the House of Commons the Digital Charter Implementation Act, or Bill C-11. But by the time this year's election was called, the bill had only passed first reading 'I don't think it would be politically tenable to go through the next session of Parliament without a reintroduction of privacy law-reform legislation,' he said. 'The question will be whether it will look like C-11'").

[290] "An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts," *House of Commons Debates*, 43-2, 150, no 30 (17 November 2020) at 1963 ("Hon. Navdeep Bains (Minister of Innovation, Science and Industry, Lib.) moved for leave to introduce Bill C-11, An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts").

proposed legislation weakens PIPEDA's conception of informed consent,[291] fails to address the reality of data de-identification,[292] and represents a "step-back overall from our current law."[293] With the dissolution of Parliament in advance of the 44th General Election, Bill C–11 died on the Order Paper, having not progressed past second reading.[294] In the 2021 federal election campaign, the Liberal Party of Canada committed to reintroducing privacy law reform in a renewed mandate.[295] Shortly after the government's re-election in September 2021, the Privacy Commissioner urged the Trudeau government to meet this promise and introduce revised privacy law reform in the new Parliament.[296]

---

[291] Teresa Scassa, "The Gutting of Consent in Bill C-11" (21 December 2021), online: <http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=336:the-gutting-of-consent-in-bill-c-11&Itemid=80> ("It sounds good until you realize that none of this is actually particularly new. Yes, the law has been tightened up a bit around implied consent and the overall wording has been tweaked. But the basic principles are substantially the same as those in PIPEDA…What has changed – and ever so much for the worse – are the exceptions to consent, particularly the ones found in sections 18 to 21 of Bill C-11").

[292] Lisa Austin & David Lie, "Bill C-11 and exceptions to consent for de-identified personal information," Schwartz Reisman Institute for Technology and Society (11 January 2021), online: <https://srinstitute.utoronto.ca/news/austin-lie-deidentified-personal-information-c11> ("It has always been problematic to try to draw a bright line between what is identifiable and what is not identifiable for the purposes of determining what is regulated and what is not. A large body of research now tells us that there is no such line, just a variety of methods to reduce the risk of re-identification and a lot of skepticism regarding eliminating this risk (for a deep dive on this issue, stay tuned for an upcoming blog post). We cannot have a regulatory architecture premised on a binary classification (identifiable/not-identifiable) if what it regulates is a spectrum of risk").

[293] Daniel Therrien, "Submission of the Office of the Privacy Commissioner of Canada on Bill C-11, the *Digital Charter Implementation Act*, 2020" (11 May 2021), online: <https://www.priv.gc.ca/en/opc-actions-and-decisions/submissions-to-consultations/sub_ethi_c11_2105/#toc2-1> ("We agree that a modern law should both achieve better privacy protection and encourage responsible economic activity, which, in a digital age, relies on the collection and analysis of personal information. However, despite its ambitious goals, our view is that in its current state, the Bill would represent a step back overall for privacy protection") [Therrien].

[294] See Brenda McPhail, "Bill C-11 was the gift that needed returning" *the Hill Times* (4 October 2021), online: <https://www.hilltimes.com/2021/10/27/bill-c-11-was-the-gift-that-needed-returning/324436> ("But that old cliché, the devil is in the details, is never truer than when applied to a 122-page piece of legislation, which died on the Order Paper when Parliament was dissolved in August").

[295] Liberal Party of Canada, *Forward. For Everyone* (Ottawa: Liberal Party of Canada, 2021), online: <https://liberal.ca/wp-content/uploads/sites/292/2021/09/Platform-Forward-For-Everyone.pdf> at 24 ("A digital society must be built on a foundation of trust. In 2019, we launched Canada's Digital Charter, which lays out 10 principles to build that foundation of trust. In November 2020, we proposed legislation to implement the Charter. We will move forward on legislation that will implement the Digital Charter, strengthen privacy protections for consumers, and provide a clear set of rules that ensure fair competition in the online marketplace").

[296] See Daniel Therrien, "Building back better requires strong, effective regulation of digital world," *the Hill Times* (27 September 2021), online: <https://www.hilltimes.com/2021/09/27/building-back-better-requires-strong-effective-regulation-of-digital-world/319480> ("Canada's federal Parliament should follow the lead of its two most populous provinces with the largest economies, Ontario and Quebec, which have made proposals towards responsible digital innovation within a legal framework that recognizes privacy as a fundamental human right").

Bill C–11 would have created a right to explanation modeled clearly on what appears in the GDPR. But in the short time Bill C–11 was being formally considered, explanation got little attention in the House of Commons.[297] Civil society likewise did not much focus on the new explanatory requirements Bill C–11 would have generated for purveyors of automated decision-making.[298] Unlike rights to explanation in the GDPR and Bill 64, and notable for our purposes here, Bill C–11 would have created a right that appears to apply even where automated processing is not the sole mechanism for decision-making. Below is the relevant provision of Bill C–11:

> If the organization has used an automated decision system to make a prediction, recommendation or decision about the individual, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision and of how the personal information that was used to make the prediction, recommendation or decision was obtained.[299]

Aside from not being limited to decision-making based solely (or exclusively in the language of Bill 64) on automated processing, this right to explanation would have notably differed from those we encountered above in at least two respects. First, Bill C–11's right to explanation depends on an individual requesting the explanation to which they are entitled. Section 63(4) would have specified that a request under this provision would have needed to be made in writing.[300] Second, Bill C–11's conception of the right to explanation has not only automated decision-making in its scope, but automated prediction and recommendation as well. It is unclear if this more expansive drafting would have had any effect on the application of the right in Bill C–11 as compared with those in Bill 64 or

---

[297] In debate held at second reading, for example, only one member, Nathaniel Erskine-Smith (Liberal, Beeches-East York), explicitly referenced the Bill's creation of a right to explanation. See e.g "An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts," *House of Commons Debates*, 43-2, 150, no 35 (24 November 2020) at 1215 ("I would like to see us go beyond algorithmic explainability to some kind of algorithmic accountability").

[298] But see Miles Kenyon, "Bill C–11 Explained" (Toronto: Munk School, 2021), online: <https://citizenlab.ca/2021/04/bill-c-11-explained/> ("Our research has shown, time and again, that organizations do not clearly explain how they collect, process, or disclose personal information under the current regime; the proposed legislation will do little to nothing to change the current abysmal state of affairs").

[299] Bill C-11, *supra* note 245, s 63(3).

[300] Bill C-11, *ibid*, s 63(4).

the GDPR. As in the case of these already-adopted rights to explanation, Bill C–11 does not set an especially clear standard with respect to what is required when an individual is subject to automated decision-making. This is a point raised by the Privacy Commissioner in his submissions to Parliament following Bill C–11's introduction. In particular, the Commissioner recommends amending section 63(3) of the Bill to "ensure the meaningfulness" of the required explanations.[301] Whether an explanation is meaningful depends on the capacity of "individuals to understand decisions made about them and facilitate the exercise of other rights such as to correct erroneous personal information."[302] This is not, in the view of the Commissioner, what Bill C–11 would have provided.[303] Instead, the Commissioner proposes that the right to explanation should specifically entitle individuals to two things, first, an accounting the nature of the decision they are subject to, including the personal information on which the decision was based, and second, detail about the "rules that define the processing and the decision's principal characteristics."[304] A right to explanation should be further complimented by a right to "contest automated decisions," roughly the capacity to request that a human review an automated decision.[305] This right to contest AI stands as an even more recent

[301] Therrien, *supra* note 319 ("First, while we acknowledge the addition in s. 63(3) of the CPPA of a new right to an explanation in relation to automated decision-making, we recommend that amendments must be made to ensure the meaningfulness of that explanation").

[302] Therrien, *ibid* ("The right to a meaningful explanation relates to existing principles for the protection of personal information, namely accuracy, openness, and individual access. This right, provided for in section 63(3) of the CPPA, should aim to allow individuals to understand decisions made about them and facilitate the exercise of other rights such as to correct erroneous personal information, including inferences. At least that is the goal of Article 15(1)(h) of the GDPR, which requires data controllers to provide individuals with "meaningful information about the logic involved" in decisions").

[303] Therrien, *ibid* ("The current obligation under section 63(3) does not provide consumers with the right to a meaningful explanation. It provides the right to know the prediction or decision, and the provenance of the information upon which this was based, but not the relationship between the personal information and the decision, nor even the elements of personal information relevant to the decision. Without the latter two elements, or at least the nature and elements of the decision to which they are being subject, or the rules that define the processing and the decision's principal characteristics, the explanation cannot be meaningful").

[304] Therrien, *ibid* ("That a standard for the level of explanation required under subsection 63(3) be enhanced to allow individuals to understand: (i) the nature of the decision they are subject to and the relevant personal information relied upon, and (ii) the rules that define the processing and the decision's principal characteristics").

[305] Therrien, *ibid* ("Additionally, individuals should be provided with a right to contest automated decisions…This right would apply both to those scenarios where an individual has provided consent for the processing of their personal

development than the right to explanation, with scholars and policymakers only recently beginning to explore the right to contestation that might be impliedly built into a statutory right to explanation.[306]

This is a consideration to which I will return in the section below. In this first part of the chapter, I have tried to give an overview of the right to explanation as it has been enacted in Europe and Canada. Each of these approaches, however conceptually distinct have much in common. As a matter of principle, these rights to explanation fundamentally aim to guarantee that individuals subject to automated decision-making have the capacity to understand why and in roughly what manner a decision affecting their interests was made. As I expressed above, this right is principally aimed at enhancing accountability in AI decision-making, to make decisions more fair, more transparent, and more reliable. In the next part, I consider what this might tangibly mean for medicine.

2. *What a Right to Explanation Might Mean for Medicine*

For persons subject to automated decision-making, the existence of a right to explanation in principle provides the capacity to obtain from the decision-maker an accounting of the reasons or factors that led to the decision. As we saw above, it is not always clear that the law contains a right to explanation. And even where a right exists, it can sometimes be difficult to understand what, precisely, it requires. This part considers what a right to explanation might mean for the practice and regulation of medicine considering the challenges I outlined in the second chapter above. This discussion will be, in parts, inevitably speculative, for the law's expectations with respect to explanation in medicine remain largely unsettled. I approach this task in two steps. First, I briefly describe how the right to explanation

information as well as those where an exception to consent was used by the organization. It serves as a complement to the right to explanation").

[306] See OECD, *Recommendation of the Council on Artificial Intelligence*, *supra* note 15 ("AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: … iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision").

is likely to be discharged in practice by decision-makers. Second, I suggest how the right to explanation might clarify the obligations of health professionals in their use of unexplainable AI.

a. Right to explanation in practice

It is one thing to require that a decision-maker explain themselves and quite another for the decision-maker to actually do so. As I outlined in the first chapter above, a not insignificant subset of AI used in medicine is deeply unexplainable. For these kinds of models, not even an initial programmer would be capable of disentangling the computational processes that produce a specific decision. It is on its face unclear how a statutory directive to explain the decisions produced by this kind of model would work. Broadly speaking, unexplainable AI users bound by a right to explanation have two options: they might interpret the right to explanation to imply that deeply unexplainable models simply cannot be used in compliance with the law, or they might find some way of engineering an explanation that would satisfy the law's demands. I will have more to say on the first of these options below. As for the second, an entire branch of AI science has in recent years taken on the task of producing explanations of the functioning of otherwise unexplainable models.[307] There are essentially two ways to approach explanation from a computing perspective. First, models can be designed prospectively to ensure that they are inherently explainable.[308] But readily interpretable models, for which a human reviewer can straightforwardly understand how decisions are made, how the relevant internal computing functions, and what kinds of factors weighed on the decision, do not respond as such to

[307] See Hagras, *supra* note 90 at 29 ("The concept of explainability sits at the intersection of several areas of active research in AI… An XAI or *transparent* AI or *interpretable* AI is an AI in which the actions can be easily understood and analyzed by humans").

[308] See Kevin Bauer et al, "Expl(AI)n It to Me – Explainable AI and Information Systems Research" (2021) 63 Business & Information Systems Engineering 79 ("Research on intrinsic interpretability methods focuses on the development of models that are inherently self-explanatory and provide an immediate human-readable interpretation about how they transform certain inputs into outputs due to their structure. Logistic regressions and decision trees are examples of simple machine learning models that are intrinsically interpretable as humans can infer their inner logic from respectively examining regressor coefficients and logic classification conditions").

the problem of deeply unexplainable AI. This approach involves developing a different kind of model, not overlaying structured reasons onto a model that otherwise could not have been explained.

A second approach is to develop *ex post* programming that uses computer science techniques to, in a manner of speaking, reverse engineer explanation.[309] One way of doing this is to train two models to operate in parallel. One is a conventional unexplainable model applied to the relevant problem or decision-making task. This is the 'live model.' A second roughly approximates the decision-making capacities of its more complex counterpart but is intentionally constructed to be simple enough to permit human review.[310] This is the 'interpretation model.' One weakness in this approach is that it is very difficult to know whether the interpretation model is faithfully representing the functioning of the live model. And to the degree that the live model is deeply unexplainable, it is impossible to know for sure.[311] Another option for *ex post* explanation is heat mapping, in which a model is designed to visually illustrate which of the factors considered in its decision-making process weighed most heavily on the outcome.[312] This is an approach particularly well suited to medical imaging, in which a model identifying which segments of an X-Ray or CT image were most salient in the model's decision-making might permit physician reviewers to better understand why the model

[309] See Blen E Kenni et al, "Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles" (2019) 7 Institute of Electrical and Electronics Engineers Access 17001 at 17003 ("in this paper, an interpretable model (i.e., fuzzy logic) is combined with a hierarchical learning method (i.e., Artificial Neural Networks). Then utilized a model induction method (i.e., reverse engineering black-box model) by applying a hybrid system to make the explanation more effective").

[310] Arun Rai, "Explainable AI: from black box to glass box" (2020) 48 Journal of the Academy of Marketing Science 137 at 138 ("Addressing the trade-off between prediction and explanation associated with deep learning models, there have been significant recent advances in post-hoc interpretability techniques—these techniques approximate deep-learning black-box models with simpler interpretable models that can be inspected to explain the black-box models. These techniques are referred to XAI as they turn black-box models into glass-box models and are receiving tremendous attention as they offer a way to pursue both prediction accuracy and interpretability objectives with AI applications").

[311] See Johan Ordish & Alison Hall, *Black Box Medicine and Transparency: Interpretable Machine Learning* (Cambridge: PHG Foundation, 2020) at 13 ("Post hoc explainers often approximate the underlying machine learning model to explain its contents. Since these explainers estimate the underlying model they may provide inaccurate answers, especially if these explainers are highly localised and taken outside their local context").

[312] See Ghassemi et al, *supra* note 93 at 746 ("Heat maps (or saliency maps)17–19 highlight how much each region of the image contributed to a given decision").

reached the decision that it did.[313] As in the parallel models case, heat mapping is limited in what it might reveal about a model's internal functioning. A model identifying which areas of a medical image most contributed to a decision, for example, does not reveal "exactly what it was in that area that the model considered useful."[314] This functionally leaves it up to the human reviewers to interpret what the heat map's meaning. And human reviewers are likely to assume that the model interpreted the image as they would have done: "You could have many explanations for what a complex model is doing. Do you just pick the one you 'want' to be correct?"[315] I will return to this problem below.

b. Right to explanation in medicine

In light of these technical approaches to achieving explanation for otherwise unexplainable AI, I turn in this section to considering how explanation might broadly function in medicine. These comments are necessarily brief. From our present vantage, with explanation rights having only recently been enacted in Canada and with little judicial contemplation of the right's operation in Europe, it is impossible to know with certainty what explanation will tangibly mean for the practice and regulation of medicine. In the straightforward sense that rights to explanation require entities engaged in automated decision-making to provide individuals an accounting of the reasons and factors implicated in a decision that affects them,[316] their application in medicine will require communication of this sort

[313] Ordish & Hall, *supra* note 334 at 13 ("Visualisation may be one of the only efficient means to meaningfully convey some of the model's function…Semantic maps (heat maps) can graphically demonstrate what an image classification found significant or indicate how it segmented the image. For example, a heat map applied to a model to identify fracture in x-ray images would highlight those elements of the image that the model found to be indicative of a fracture").

[314] Ghassemi et al, *supra* note 93 at 746 ("Even the hottest parts of the map contain both useful and non-useful information (from the perspective of a human expert), and simply localising the region does not reveal exactly what it was in that area that the model considered useful. The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease").

[315] Aaron M Bornstein, "Is Artificial Intelligence Permanently Inscrutable?" (2016) 40 Nautalis, *quoting* Cynthia Rudin, online: <https://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable> ("Rudin's concerns echo the famous dictum that there may be no simpler model of the visual system than the visual system itself. 'You could have many explanations for what a complex model is doing,' she says. 'Do you just pick the one you 'want' to be correct?'").

[316] See Bill 64, *supra* note 16, s 102 ("…of the reasons and the principal factors and parameters that led to the decision").

of information to patients. While certain scholars suggest that explanation will assist in alleviating the conceptual difficulty generated by unexplainable AI that I outlined in the second chapter above,[317] there is as yet uncertainty about how a right to explanation will affect medical decision-making.

On first brush, for example, Quebec's Bill 64 applies to "any person carrying on an enterprise [that] uses personal information to render a decision based exclusively on an automated processing."[318] There are at least two potential sources of unclarity in this legislative statement. First, whether clinicians, for the purposes of this provision, are "carrying on an enterprise" does not admit of a simple assessment. In the *Civil Code*, for example, enterprise is defined broadly as the "carrying on by one or more persons of an organized economic activity, whether or not it is commercial in nature."[319] In this framing, a clinician's economic and employment relationships might be materially relevant in determining who precisely is obliged by Bill 64 to communicate an explanation of an automated decision. Under certain conditions, the enterprise captured by Bill 64's right to explanation may be hospital or clinic, while, in others, an individual clinician may be expected to furnish an explanation. I do not engage in a detailed assessment of this issue here, which is far beyond our present scope. I rather flag this formulation as a potential source of uncertainty about how explanation in medicine will be addressed. Second, application of Bill 64's right to explanation is specifically limited to decisions "based exclusively on automated processing." While there may come a time, perhaps not so long from now, where certain kinds of medical decisions are deferred in their entirety to AI models, most clinicians are in the medium-term likely to draw on automated decision-making

[317] See e.g. Sullivan & Schweikart, *supra* note 14 at 164 ("The rise of black-box AI and its use in medicine complicates application of existing tort law when trying to resolve claims of malpractice. If a patient becomes injured by use of an AI technology (black-box AI in particular), current legal models are insufficient to address the realities of these innovations. New legal solutions that craft novel legal standards and models that address the nature of AI, such as AI personhood or common enterprise liability, are necessary to have a fair and predictable legal doctrine for AI-related medical malpractice").

[318] Bill 64, *supra* note 16, s 102.

[319] See art 1525 CCQ ("The carrying on by one or more persons of an organized economic activity, whether or not it is commercial in nature, consisting of producing, administering or alienating property, or providing a service, constitutes the operation of an enterprise").

as part of a broader clinical approach that involves human oversight and review. Clinical decision-making, in other words, is probably going to be based only partially on automated processing. But courts might nevertheless interpret certain highly particularized fact patterns to satisfy this exclusive automated processing clause. Should a physician fail to sufficiently review or monitor an automated decision-making process, for example, it is not inconceivable that a court could find the existence of a right to explanation. One final point is worth noting. Bill 64's right to explanation is not automatically triggered but is rather activated on a request by the individual subject to an automated decision.[320] It might end up being quite administratively difficult for patients to submit, and for health systems to review, requests for explanation.[321] How each of these factors will affect the execution of rights to explanation in medicine is, I think, highly uncertain.

More broadly speaking, and assuming that explanation can be effectively implemented in medicine, the right conveyed in Bill 64 and elsewhere might predictably have three general kinds of effects on medicine. First, greater access to model explanations might change the way clinicians behave, particularly in their decisions whether and how to use unexplainable AI in medical practice. One possibility is that clinicians will generally opt only to use models for which explanations are readily available, or for which explanations are aligned with prevailing clinical judgement. It could be that a right to explanation will generally create regulatory pressure on the developers of medical imaging models to produce explanatory programming. In the event courts interpret enacted rights to explanation broadly, developers may be effectively required to at least have the capacity to provide minimal reasons supporting the generation of an automated decision. These effects might further

[320] Bill 64, *supra* note 16, s 102 ("He must also inform the person concerned, *at the latter's request*…of the reasons and the principal factors and parameters that led to the decision").

[321] Canada's publicly funded Medicare systems, in which triage and carefully targeted funding allocation play an especially important role in keeping costs low and care accessible, might be particularly prone to administrative complexity in the creation of an implementation framework for rights to explanation. See generally Julie Fiset-Laniel, Ak'ingabe Guyon, Robert Perreault & Erin C Strumpf, "Public health investments: neglect or wilful omission? Historical trends in Quebec and implications for Canada" (2020) 111 Canadian Journal of Public Health 383.

address some of the unclarity surrounding the obligations of clinicians in the use of unexplainable AI. To the degree clinicians roughly understand how a model operates, for example, both reviewing courts and the professionals themselves will be better able to assess which models may be trusted and which ought to be excluded from having a role in medical practice. A right to explanation, in other words, may assist the law in understanding and assigning the standard of care for using unexplainable AI. Second, and as a sort of corollary of the above, a right to explanation might clarify the chain of causation leading to patient injury.[322] Requiring model developers to provide an accounting of the way particular decisions are made might straightforwardly lend itself to a better understanding of the causal chain producing injury in medicine. Third, drawing together the points above, an effectively functioning right to explanation might permit easier access to redress for patients injured when their care involves the use of an unexplainable AI. Explanations, after all, occupy a dominant role in legal systems and the unexplainable character of certain models is likely to greatly complicate the ability of injured persons to make out a case against their injurer.[323]

But to the extent that explanations generated in contemplation of the statutory frameworks described above shed some light on the law's expectations, the practical capacity of individuals to make out a malpractice case might be enhanced. Recall that in the *Ola* case, a court ordered disclosure of the personal data used in the training of its unexplainable model on the request of the individuals subject to its decision-making processes.[324] While this is notably different than the kinds of programming-derived explanations I described in the part above, it provides at least a minimal

[322] See Hagras, *supra* note 90 at 30 ("This will allow both users and stake-holders to understand the AI's cognition and empower them to determine when to trust or distrust the AI. This establishes the ability to satisfy the abovementioned points of transparency and causality and address the system bias, fairness, and safety").

[323] Katie Atkinson, Trevor Bench-Capon & Danushka Bollegala, "Explanation in AI and law: Past, present and future" (2020) 289 Artificial Intelligence 1 at 18 ("Explanation is an essential feature of legal systems intended to predict case outcomes and so it is crucial that this aspect be developed for machine learning systems intended for deployment on such tasks").

[324] *Ola Netherlands BV*, *supra* note 261 at para 4.62 ("Ola must provide access to the (personal) data referred to above under 4.25, 4.45, 4.47, 4.49 and 4.52").

framework in which triers of fact may assess a claim for redress. And yet processed data or underlying model code probably does not provide clinicians or non-expert patients the kind of accounting for a model's functions that would entirely clarify the kinds of considerations described here. In the event courts interpret rights to explanation to require only the disclosure of framework data, rather than programming-derived explanations such as those conveyed in heat mapping or parallel modeling, then the right's effects on malpractice may well be exceptionally minor.

To be sure, the effects of rights to explanation on medicine will not occur in a regulatory vacuum. Those rights that have thus far been implemented generally exist as but one part of sweeping privacy and data management statutory frameworks. It may be that the most important controls on automated decision-making stem not from explanation but from other legislative implements, such as rights of access, disclosure, or erasure.[325] One approach of interest may be that taken by France in its *Code de la santé publique*. There, the legislator creates an obligation on the part of health professionals to ensure that individuals subject to automated decision-making for the purposes of prevention, diagnosis, or care are informed of the clinician's use of an algorithm.[326] Bill 64 conveys a similar obligation.[327] This approach might have several significant advantages over a right to explanation. For one thing, the scope of the obligation is clearly delimited and is specifically directed at the medical context. For another thing, the obligation is relatively easily discharged, consisting

[325] See e.g. Tiago Sérgio Cabra, "Forgetful Al: Al and the Right to Erasure under the GDPR" (2020) 3 European Data Protection Law Review 378 at 388 ("It is not our belief that the GDPR and Al development cannot coexist. However, both developers of AI-enabled technology and companies using Al will have to undertake a significant effort to ensure that these technologies are used in a manner that is GDPR-compliant. Then, how to avoid disruptions of databases and a potential chilling effect on Al caused by the right to erasure? For data controllers the advice is to take into account the principles of privacy by design and privacy by default and try to build algorithms that are resistant to the erasure of certain data entries").

[326] *Code de la santé publique*, JO, 2 août 2021 (NC), art L4001-3 ("Le professionnel de santé qui décide d'utiliser, pour un acte de prévention, de diagnostic ou de soin, un dispositif médical comportant un traitement de données algorithmique dont l'apprentissage a été réalisé à partir de données massives s'assure que la personne concernée en a été informée et qu'elle est, le cas échéant, avertie de l'interprétation qui en résulte").

[327] Bill 64, *supra* note 16, art 102 ("Any person carrying on an enterprise who uses personal information to render a decision based exclusively on an automated processing of such information must, at the time of or before the decision, *inform the person concerned accordingly*").

only in conveying to patients information that is well within the clinician's capacity to convey. But notably, a duty to inform that automated processing is used in patient care does little to address the specific challenges I detailed in the second chapter above. To inform patients that unexplainable AI was used in their care, for example, does not resolve uncertainty about how clinical fault or causation ought to be assessed. And in any case, duties to inform in France and Quebec are not precisely alternatives to explanation. They are, after all, situated in broader regulatory frameworks that include the generation of explanatory rights. That there are other options for managing the use of medical AI does not dilute the urgency of addressing explanation. After all, as I suggested above, rights to explanation occupy a dominant and apparently growing position in the regulation of AI. In the next part, I argue that they should not.

3. *Why the Right to Explanation Fails*

Explanation does not address the challenges for medical practice and regulation outlined in the second chapter and probably makes things worse. There are two principal reasons for this. First, the right to explanation tries to guarantee something that cannot be achieved: a true accounting of the reasons for which automated decisions are produced. Explanatory techniques as they presently stand offer only the mirage of explanation, a facsimile that does not, and likely cannot, peer under unexplainable AI's functional hood. Second, rights to explanation fundamentally depend on an unfounded assumption that human oversight is capable of tempering AI's worst effects. There is no compelling reason, I argue, to think that human review is useful in the control of unexplainable AI. Humans will instead likely limit the utility of AI in medicine while also replicating its harmful impacts. I suggest that rights to explanation for automated decision-making promote an untenable attitude of exceptionalism about an unfamiliar but otherwise not conceptually distinct mode of medical intervention. By thinking of unexplainable AI as essentially distinct and by applying a bespoke regulatory theory to its control,

rights to explanation work to deprive patients of potentially improved outcomes while also distracting from sources of significant harm prevalent across medicine. I present each of these views in turn.

a. Explanations do not explain

Hard as they might try, rights to explanation do little to actually illuminate how unexplainable models function. Commonly applied *ex post* explanatory methods do not so much represent how automated decision-making really works as approximate the kinds of factors that may have played some role in an otherwise indefinite process.[328] As I hinted at above, techniques for producing *ex post* explanations encounter several challenges. One problem is that *ex post* explanations, particularly those produced by parallel programming, merely approximate decision-making processes taken by an unexplainable model. These kinds of explanations are necessarily estimative and might, in consequence, sometimes end up being wrong.[329] This might be especially worrying where techniques for *ex post* explanation produce locally interpretable explanations that apply only in specific contexts.[330] These kinds of explanations, taken to apply in a more general context than that for which they were developed, will likely not reflect explanatory fidelity, that is, will likely not faithfully represent the function of the decision-making model.[331] Another significant problem is that *ex post* explanations often give only a

[328] See e,g, Ghassemi et al, *supra* note 93 at 746 ("The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease").

[329] Ordish & Hall, *supra* note 334 at 14 ("Post hoc explainers often approximate the underlying machine learning model to explain its contents. Since these explainers estimate the underlying model they may provide inaccurate answers, especially if these explainers are highly localised and taken outside their local context").

[330] See Riccardo Guidotti et al, "A Survey of Methods for Explaining Black Box Models" (2018) 51:5 ACM Computing Surveys 93:1 at 93:14 ("Given a black box and an input instance, the outcome explanation problem consists in providing an explanation for the outcome of the black box on that instance. It is not required to explain the whole logic underlying the black box but only the reason for the prediction on a specific input instance. We formalize this problem by assuming that first an interpretable local model cl is built from the black box b and the instance x, and then an explanation is derived from cl").

[331] Ordish & Hall, *supra* note 334 at 21 ("Post hoc explainers also have weaknesses and risks. Notably, in terms of fidelity, these explainers are approximations").

partial account of decision-making processes.[332] This is evident in heat mapping for medical imaging models. As I outlined above, these techniques only indicate which areas of an image a model identifies as salient. But localizing image regions relevant in a decision-making process does not on its own uncover how the regions are considered, whether, for example, a salient region contains irrelevant sub-regions.[333] In respect of a specific model, heat mapping tells us little more than "where the [model] is looking within the image."[334] It does not reveal what the model is doing with the data.[335]

One implication of this is that heat mapping methods cannot generally reveal the kinds of factors a model considers with any kind of precision. Imaging models might focus on the right regions of a scan, but might consider image quality factors that no human reviewer would think has anything to do with an underlying disease.[336] Unexplainable models, after all, will sometimes make decisions in ways humans find surprising or counterintuitive. When this happens, human reviewers might tend toward confirmation bias, substituting programming-generated analysis for their own intuition. It is widely understood that humans are highly susceptible to a multiplicity of cognitive errors, one of the most powerful of which is confirmation bias, where evidence that conflicts with an individual's held

[332] See Rudin, *supra* note 94 at 208 ("Even if both models are correct (the original black box is correct in its prediction and the explanation model is correct in its approximation of the black box's prediction), it is possible that the explanation leaves out so much information that it makes no sense").

[333] Ghassemi et al, *supra* note 93 at 746 ("Even the hottest parts of the map contain both useful and non-useful information (from the perspective of a human expert), and simply localising the region does not reveal exactly what it was in that area that the model considered useful").

[334] Rudin, *supra* note 94 at 208 ("Saliency maps can be useful to determine what part of the image is being omitted by the classifier, but this leaves out all information about how relevant information is being used. Knowing where the network is looking within the image does not tell the user what it is doing with that part of the image, as illustrated in Fig. 2. In fact, the saliency maps for multiple classes could be essentially the same; in that case, the explanation for why the image might contain a Siberian husky would be the same as the explanation for why the image might contain a transverse flute").

[335] See Ordish & Hall, *supra* note 334 at 21 ("However, knowing where the model is looking does not tell us what the model is doing with that part of the image").

[336] Ghassemi et al, *supra* note 93 at 746 ("The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease").

views is disregarded.[337] Considering that even our best *ex post* explanatory methods leave significant room for interpretation, it may be that human reviewers will interpret machine-generated explanations to align with their preconceptions, effectively substituting their own intuition for the unexplainable factors that actually produced the decision.[338] One way this might manifest is in excessive deference to a model's judgement.[339] Humans might, for example, ascribe positive interpretations to otherwise ambiguous explanations, finding meaning that aligns with human intuition even where the model's actual approach is entirely unknowable. This is best illustrated in the medical imaging context. Heat map explanations of a pulmonary scan might identify certain salient structures as the dispositive elements in a model-generated prognosis. And human reviewers might agree that those structures identified on the heat map, for various complicated medical reasons, confirm the model's judgement. But we have no way of knowing whether the model made its decision on factors no human would consider medically relevant, such as pixel density or texture.[340] What is important here is that the essential ambiguity of machine-generated explanations for automated decision-making means that human beings are left to fill in the explanatory gaps. Human reviewers will substitute interpretations derived from their own judgement, subject as it is to cognitive error and bias. Explanation in this

[337] Max Rollwage et al, "Confidence drives a neural confirmation bias" (2020) 11:2634 Nature Communications 1 at 2 ("This polarization is most evident when opposing parties are highly confident in their positions. A psychological-level explanation for such entrenchment is the idea that people selectively incorporate evidence in line with their beliefs, known as confirmation bias").

[338] Robert Challen et al, "Artificial intelligence, bias and clinical safety" (2019) 28 BMJ Quality & Safety 231 at 234 ("As humans, clinicians are susceptible to a range of cognitive biases which influence their ability to make accurate decisions. Particularly relevant is 'confirmation bias' in which clinicians give excessive significance to evidence which supports their presumed diagnosis and ignore evidence which refutes it. Automation bias describes the phenomenon whereby clinicians accept the guidance of an automated system and cease searching for confirmatory evidence").

[339] See Raymond R Bond et al, "Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms" (2018) 51 Journal of Electrocardiology 6 at 7 ("Nevertheless, clinicians can be overly influenced by the [automated diagnosis] which can be referred to as automation bias [13]. Automation bias exists when humans over rely on automation to complete a task. This phenomenon is similar to other cognitive biases such as anchoring and confirmation bias").

[340] Ghassemi et al, *supra* note 93 at 746 ("The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease").

context then, does not so much explain what automated decision-making processes in reality does, but rather reflects whatever a model's human interpreter assumes a model to have done. If this is right, then the statutory rights to explanation I outlined above do little to provide individuals what the law intends. Rights to explanation cannot be expected to provide recourse toward understanding how an unexplainable model works. They provide at best an illusion of explanation.

b. Human oversight is not effective

Even if rights to explanation were able to provide what they intend, accurate and meaningful accounts of the reasons and principal factors considered in automated decision-making, they would still depend on a debateable assumption that human oversight of AI is effective and valuable. It is neither. Debate surrounding the right to explanation is premised fundamentally on the view that human reviewers ought to play an important role in the control and oversight of AI.[341] This is sometimes called human-in-the-loop decision-making,[342] which has in recent years become an influential principle in AI ethics principles and guidance documents.[343] It is on its face reasonable that human oversight of AI should be a dominant concern for ethicists and policymakers. Our best AI models, after all, are at best poorly understood black boxes entrusted to wield significant influence over some of the most important

[341] See Bryan Casey, Ashkon Farhangi & Roland Vogl, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise" (2019) 34 Berkeley Technology Law Journal 143 at 170 ("The A29WP's guidance on automated decision-making included numerous provisions intended to clarify the 'right to explanation'-stemming from a collection of rights that the A29WP referred to as the rights 'to be informed,' 'to obtain human intervention,' and 'to challenge [a] decision' made by certain automated systems").

[342] See Therese Enarsson, Lena Enqvist & Markus Naarttijärvi, "Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts" (2021) Information & Communications Technology Law 1 at 2 ("The impetus to implement hybrid decision-making may vary. In some cases, it may be driven by ambitions of increased efficiency where reducing human discretion is a specific goal which cannot fully be realized due to technical or legal constraints. In other areas, such as online moderation, the need for human contextual analysis is well known, but the sheer scope of the task facing moderators and external pressures calls for further automation. However, in many cases, keeping a *human in the loop* is a deliberate attempt to maintain human agency and accountability, and to provide legal safeguards and quality control").

[343] See Montreal Declaration, *supra* note 196 at principle 4 at principle 4 ("[AI systems] should not be implemented to replace people in duties that require quality human relationships, but should be developed to facilitate these relationships"); UNESCO, *First Draft of the Recommendation on the Ethics of Artificial Intelligence*, SHS/BIO/AHEG-AI/2020/4 REV.2, Paris (2020) at s 35 ("It must always be possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems to physical persons or to existing legal entities. Human oversight refers thus not only to individual human oversight, but to public oversight, as appropriate").

aspects of our lives. To cede decision-making power to unexplainable machines feels like asking for trouble. This is especially true when the machines in question end up perpetuating discrimination,[344] threatening privacy,[345] or depriving individuals of their right to procedural fairness.[346] Both rights to explanation and human oversight are sometimes promoted as a way of warding against many of these harms.[347] But human oversight offers at best a thin layer of protection against the risks of automated decision-making. I suggested above that even our best explanations of automated decision-making are essentially rough approximations, and that human intuition serves as an "unacknowledged bridge" between what an explanation describes and a normative evaluation about whether the model operated reasonably.[348] While much attention is given, for example, to the possibility that AI will generate biased outcomes, considerably less is given to the possibility that human oversight, influenced as it is by deference to automation and error-prone intuition, may be a profound source of bias as well.[349]

---

[344] See Sandra Wachter, Brent Mittelstadt & Chris Russell, "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI" (2021) 41 Computer Law & Security Review 1 at 5 ("Artificial intelligence creates new challenges for establishing prima facie discrimination. By definition claimants must experience or anticipate inequality. Compared to traditional forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, and intangible. These characteristics make it difficult to detect and prove as victims may never realise they have been disadvantaged").

[345] See e.g. Elyse Tom et al, "Protecting Data Privacy in the Age of AI-Enabled Ophthalmology" (2020) 9:2 Translational Vision Science & Technology 1 at 2 ("The combination of Big Data and AI also offers many potential benefits for health-care systems, including increased productivity with decreased costs, as well as reductions in medical error. New data privacy problems have arisen with the use of this technology, however, leading to concerns about the balance between innovation and privacy and the need for better data protection methods that can evolve along with Big Data and AI").

[346] See Jennifer Raso, "Administrative Law" in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 181–202 at 190 ("To date, no reported cases exist in which an algorithmically-driven decision has been challenged on procedural fairness grounds. Nonetheless, such decisions likely violate notice and disclosure requirements, and may raise concerns about whether an adequate hearing or reasons were provided").

[347] See Mariarosaria Taddeo & Luciano Floridi, "How AI can be a force for good" (2018) 361:6404 Science 751 ("The case of COMPAS, an AI legal system that discriminated against African-American and Hispanic men when making decisions about granting parole, has become infamous. Robust procedures for human oversight are needed to minimize such unintended consequences and redress any unfair impacts of AI").

[348] Selbst & Barocas, *supra* note 92 at 1086 ("In most cases, intuition serves as the unacknowledged bridge between a descriptive account and a normative evaluation").

[349] See Ghassemi et al, *supra* note 93 at 748 ("In the example of heat maps, the important question for users trying to understand an individual decision is not where the model was looking but instead whether it was reasonable that the model was looking in this region. By conflating these questions and allowing intuition to bridge the gap, there is a serious risk of introducing harmful biases into decision making").

There is some irony, I think, in the view that human reviewers will be generally capable of detecting biased decision-making in AI. Automated processes, after all, are not inherently directed toward discrimination. AI does not create bias so much as replicate social and systemic features that are fundamentally the creations of human beings. To think of human review as a solution to this problem is, in some sense, to leave the fox in command of the henhouse. This is not to say that human reviewers will not on occasion be quite effective in the control of biased AI decision-making. But our present capacities to interrogate AI explanation suggest that this is unlikely to be a successful approach in aggregate. Not only are human reviewers prone to precisely the same biased tendencies as AI (the models learned from us, after all), but they also have the parallel habit of deferring excessively to automated decision-making. This might have the effect of making bias worse rather than better. Our vigilance against error in AI decision-making is, in virtue of our cognitive tendencies and the technical limits of explanation, highly compromised.[350] All of this being the case, human oversight facilitated by explanation will probably not be effective against AI-generated bias and may even make things worse. Human oversight of AI conceptually misdirects our attention away from the human systems actually responsible for generating bias and inequality. Instead, explanation adopts a view that human beings are a reliable safeguard against what are, at bottom, human problems. AI is not on its own a cause of inequality in medicine. Human oversight mistakenly assumes that it is.

To be sure, bias is just one of the numerous AI-associated challenges in medicine for which human oversight is expected to respond. While not the only problem raised in the use of unexplainable AI, how explanation fails to control bias helps to illustrate a broader point. Human oversight is in an important sense incapable of effectively addressing any of the serious challenges generated by the

---

[350] See Ghassemi et al, *supra* note 93 at 748 ("Left unchecked, an AI system could operationalise these biases on a large scale. It is implied that explainability could allow us to catch discriminatory behaviour more readily. Unfortunately, as outlined above, this possibility is not reflected in the current state of explainability research, and reliance on explanations might even decrease our vigilance for these behaviours").

unexplainable character of our most effective AI for the straightforward reason that explanation as such is foundationally something of an illusion. There is a related, though not identical reason for which human oversight of unexplainable AI might be conceptually incoherent. Human oversight treats AI as a technologically exceptional phenomenon requiring a legally exceptional response.[351] Assuming that much of AI is deeply unexplainable in the way described in the first chapter, it is perhaps not entirely surprising that this should be the case. In using unexplainable models, after all, we end up deferring significant decision-making discretion to an effectively mysterious process: just stating the problem appears to mitigate for a uniquely stringent approach to regulating AI. Cultural perceptions surrounding computers and automation also probably contribute to a sense of uneasiness about unexplainable decision-making. Fear that our presently limited models will one day transform into artificial general intelligence (AGI), models that not just exceed the cognitive capacities of humans in one domain, but in *all* domains,[352] may prompt a cautious approach to AI oversight. Setting aside that it is not even clear whether AGI is possible,[353] how human oversight of AI-mediated decision-making would effectively ward against its development is a complete mystery.

[351] See Jean-Christophe Bélisle-Pipon et al, "What Makes Artificial Intelligence Exceptional in Health Technology Assessment?" (2021) 4 Frontiers in Artificial Intelligence 1 at 2 ("The present article was guided by this question: what makes artificial intelligence exceptional in health technology assessment? To our knowledge, this is the first review on this topic. After describing the methodology of the review, we will provide a comprehensive overview of AI-specific challenges that need to be considered to properly address AIHTs' intrinsic and contextual peculiarities in the context of HTA. This will lead to point possible explanations of this exceptionalism and solutions for HTA").

[352] See Brian S Haney, "The Perils and Promises of Artificial General Intelligence" (2018) 45:7 Journal of Legislation 151 at 151–152 ("Further, it has often been the case that once an Al system reaches human level performance at a given task, shortly thereafter that same Al system exceeds the performance of the most skilled humans in completing that task. Many Al researchers expect that Al systems will eventually reach and then exceed human-level performance in all tasks… Even those who doubt whether Artificial General Intelligence ('AGI'), Al capable of accomplishing any goal, will be created in the future, still agree that Al will have profound implications for all domains, including: healthcare, law, and national Security").

[353] See e.g. Ragnar Fjelland, "Why general artificial intelligence will not be realized" (2020) 7:10 Humanities & Social Sciences Communications 1 at 3 ("Some spectacular breakthroughs have been used to support the claim that AGI is realizable within the next few decades, but I will show that very little has been achieved in the realization of AGI. I will then argue that it is not just a question of time, that what has not been realized sooner, will be realized later. On the contrary, I argue that the goal cannot in principle be realized, and that the project is a dead end").

More to the point, though, while there may be a sense in which human oversight of unexplainable AI is a useful regulatory approach, it is not by treating automated decision-making as a uniquely difficult regulatory challenge that such oversight is likely to succeed. Though the kind of human oversight underpinning rights to explanation imply that the unexplainable character of automated decision-making systems requires special kinds of regulation, this attitude conflicts with much of how contemporary medicine works. It is just not the case that AI's tendency to operate in unexplainable ways is an unforeseen development, particularly in medicine. Black box systems are already prevalent and deployed safely and effectively to the benefit of millions of patients. Several authors have pointed out that many of our best drugs and devices function in ways we do not fully understand.[354] Acetaminophen, as I described above, has a mysterious mechanism of action.[355] Yet acetaminophen is generally safe and effective. And we know this even if we do not know how its chemical and physiologic functions operate. Analogously, not knowing how an unexplainable model works at the level of computer programming does not by necessity imply that it is dangerous. It does not even mean that the model is more likely than other health interventions to produce biased or inequitable outcomes. Just as we know that acetaminophen works, we might confidently deduce that a rigorously and independently tested model is safe and effective. This, of course, requires human oversight in the form of regulatory control and randomized controlled trials.[356] But it is not the kind of human oversight that approaches unexplainable AI as something never seen before in medical decision-making. Human oversight that approaches AI as a medical innovation, testable and

[354] Ghassemi et al, *supra* note 93 at 748 ("The medical system is already extremely adept at evaluation and validating various kinds of black-box systems, as many drugs and devices function, in effect, as black boxes").

[355] Peter Kirkpatrick, "New clues in the acetaminophen mystery" (2005) 4 Nature Reviews Drug Discovery 883 at 883 ("Although acetaminophen (paracetamol) has been used clinically for more than a century, its mode of action is still not clear"); See also Ghassemi et al, *supra* note 93 at 748 ("An often cited example is acetaminophen, which, despite having been used for more than a century, has a mechanism of action that remains only partially understood").

[356] Ghassemi et al, *supra* note 93 at 748 ("Despite competing explanations for how acetaminophen works, we know that it is a safe and effective pain medication because it has been extensively validated in numerous randomised controlled trials (RCTs). RCTs have historically been the gold-standard way to evaluate medical interventions, and it should be no different for AI systems").

reviewable even if not entirely understandable, may be our best bet for ensuring that unexplainable models promising to improve patient care can safely and effectively be deployed. But oversight consisting in a demand for explanation is doomed to fail. Explanation asks for more than we reasonably can expect. It is a salve that cannot mend unexplainable AI's wounds.

Conclusion

This chapter gave an overview of the right to explanation as it has been enacted in the European Union and Quebec. I considered in broad strokes how rights of this kind might tangibly affect medical practice and argued that rights to explanation are fundamentally incapable of remedying conceptual challenges raised by unexplainable AI. I concluded by observing that unexplainable practices may be more common in medicine than we sometimes think. Under that view, unexplainable AI might not constitute a wholly unprecedented phenomenon.

CONCLUSION

Unexplainable AI creates for medicine a monumental but perhaps not entirely unprecedented set of challenges. This essay surveyed what these challenges might be and how prevailing regulation proposes to respond. In this essay's first chapter, I introduced artificially intelligent decision-making in the medical context. I first gave an overview of foundational AI concepts, defined unexplainable AI, and surveyed some of the deep learning models presently approved for clinical use in Canada. In the second chapter, I considered some of the ways unexplainable AI might raise conceptual challenges for the practice and regulation of medicine. I did this by inquiring into the potential effects of unexplainable medical AI on malpractice law. This second chapter suggested that unexplainable AI might have implications for the way malpractice law assess clinical fault and causation. In the third chapter, I attended to the right to explanation, perhaps the predominant policy response for controlling theorized and real-world impacts of unexplainable AI. Though designed to assuage the kinds of concerns I present in the second chapter, I argue that rights to explanation miss the mark. I did this by first summarizing the legislative operation of rights to explanation in Europe and Quebec. I briefly predicted how rights to explanation might impact medical practice and argued that they are probably ineffective. Explanation, in fact, probably ends up doing more harm than good.

I closed this essay by suggesting that unexplainable AI is perhaps not so unlike more conventional medical practices and innovations. Much of what happens in the clinic is unexplainable, including some of the drugs and devices on which we rely most extensively. Even clinical judgment, a conventionally indispensable part of patient care, may be difficult to pin down or to explain concretely.[357] None of this is usually a concern for jurists or clinicians, for medical science is capable

[357] See e.g. Tim Thornton, "Tacit knowledge as the unifying factor in evidence based medicine and clinical judgement" (2006) 1:2 Philosophy, Ethics & Humanities in Medicine 1 at 2 ("By contrast, clinical expertise is not codifiable. It depends instead on skilled judgement drawing on personal experience"); Julia Amann, Alessandro Blasimm, Effy Vayena et al, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective" (2020) 20 BMC Medical Informatics & Decision Making 1 at 7 ("Additionally, physicians will rarely have the time to perform an in-depth analysis of why their clinical judgement is in disagreement with the AI system. Thus, looking merely at a performance output is not sufficient in the clinical context").

of assessing whether interventions are safe and effective. Even if the mechanisms underlying a particular decision or intervention are difficult to robustly articulate, the system tends to work, our most reliable medical tools work as we intend them, and clinical judgement is often vindicated. This being so, we might wonder why the conceptual challenges raised in the use of unexplainable medical AI need addressing at all. If unexplainable AI is in reality not so different than the innumerable innovations made throughout the history of medicine, then surely the conceptual challenges it raises for malpractice law and medical regulation will resolve themselves in time. And while it is nearly certainly right that medical and legal practices will evolve in ways that accommodate unexplainable medical AI, it would be a mistake to therefore believe that the unexplainable character of our best medical AI need not be interrogated. It may be that there is no firm difference in kind extending between unexplainable AI and prior unexplainable medical innovation. But there is a vast difference, I think, in magnitude. AI's applications in medicine are universally expected to be widespread and unprecedented. Hardly any specialty or area of practice will be immune to AI's reach. This on its own might make the kinds of conceptual problems I identified in the second chapter worth addressing.

Beyond that, unexplainable AI models have generally not been in clinical use long enough to have established the kind of popular trust among clinicians and patients that is available to innovations like acetaminophen. We know acetaminophen works in part because humans have been using it for generations. AI's relative novelty appears to have become a source of suspicion, particularly among policymakers. No one, after all, is trying to implement a right to explanation for painkillers. I have tried to argue that AI raises genuine challenges deserving serious contemplation. But the best and most coherent way of addressing them is not with maladroit policy premised on a mythology of explanation, but with thoughtful regulatory oversight and review. Rights to explanation, I think, fundamentally misconstrue how unexplainable AI works and exaggerate the human capacity for effective oversight. There may be fields in which rights to explanation are appropriate, in which even the facsimile of an explanation is more normatively or procedurally valuable than the reality that such

explanation is essentially an illusion. In the administrative law context, for example, procedural fairness rules might mitigate heavily in favour of even highly approximative explanations of automated decision-making. But medicine does not appear to be an analogous domain. Though administrative law guarantees fairness of process, medical malpractice law is grounded principally on the promise that patients are entitled to care aligned with our best scientific evidence, as it is embodied in the practical consensus of professionals engaging in an empirically oriented craft. Of course, what the evidence supports is often quite unclear. But clinicians generally try to make the best decisions that they can, influenced as they are by institutional and normative considerations that sometimes operate to advance our worst instincts. Unexplainable AI, when it works in its most faultless form, should tend toward improving clinical judgement, making it less dependent on the kinds of contingent and irrelevant factors to which human decision-makers might be prone. Rights to explanation would thwart these aims, returning human intuition to the centre of clinical decision-making. Insofar as we think AI advances the vision of more equitable, inclusive, effective care, itself a contestable assumption, explanation ends up being entirely counterproductive.

REFERENCES

Secondary Materials

Alpaydin, Ethem, *Introduction to Machine Learning* (Cambridge: MIT Press, 2014).

Amann, Julia, Alessandro Blasimm, Effy Vayena et al, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective" (2020) 20 BMC Medical Informatics & Decision Making 1.

Atkinson, Katie, Trevor Bench-Capon & Danushka Bollegala, "Explanation in AI and law: Past, present and future" (2020) 289 Artificial Intelligence 1.

Austin, Lisa & David Lie, "Bill C-11 and exceptions to consent for de-identified personal information," Schwartz Reisman Institute for Technology and Society (11 January 2021), online: <https://srinstitute.utoronto.ca/news/austin-lie-deidentified-personal-information-c11>.

Avraham, Ronan & Max M Schazenbach, "Medical Malpractice" in Francesco Parisi, ed, *Oxford Handbook of Law & Economics, Volume II: Private & Commercial Law* (Oxford: Oxford University Press, 2017) 120–147.

Babuta, Alexander, Marion Oswald & Ardi Janjeva, "Artificial Intelligence and UK National Security Policy Considerations" (London: Royal United Services Institute for Defence and Security Studies, 2020).

Bathaee, Yavar, "The Artificial Intelligence Black Box and The Failure of Intent and Causation" 31:2 (2018) Harvard Journal of Law & Technology 889.

Baudouin, Jean-Louis, Patrice Deslauriers & Benoît Moore, *La responsabilité civile, Volume 1: Principes généraux*, 9th edition (Montreal: Éditions Yvons Blais, 2020).

Baudouin, Jean-Louis, Patrice Deslauriers & Benoît Moore, *La responsabilité civile, Volume 2: Responsabilité professionelle*, 9th edition (Montreal: Éditions Yvons Blais, 2020).

Baudouin, Louis, *Le Droit Civil de la Province de Québec: Modèle Vivant de Droit Comparé* (Montreal: Wilson & Lafleur, 1953).

Bauer, Kevin et al, "Expl(AI)n It to Me – Explainable AI and Information Systems Research" (2021) 63 Business & Information Systems Engineering 79.

Beam, Andrew L & Isaac S Kohane, "Big Data and Machine Learning in Health Care" (2018) 319:13 JAMA 1317.

Bélisle-Pipon, Jean-Christophe et al, "What Makes Artificial Intelligence Exceptional in Health Technology Assessment?" (2021) 4 Frontiers in Artificial Intelligence 1.

Benjamens, Stan, Pranavsingh Dhunnoo & Bertalan Meskó, "The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database" (2020) 3 Nature Digital Medicine 1.

Black, Lee, “Effects of Malpractice Law on the Practice of Medicine” (2007) 9:6 American Medical Association Journal of Ethics 437.

Bond, Raymond R et al, “Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms” (2018) 51 Journal of Electrocardiology 6.

Bornstein, Aaron M, “Is Artificial Intelligence Permanently Inscrutable?” (2016) 40 Nautalis, online: <https://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable>.

Bringsjord, Selmer & Naveen Sundar Govindarajulu, “Artificial Intelligence” in Edward N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2020).

Bryson, Joanna J, “The future of AI’s impact on society” (2019) MIT Technology Review.

Burrell, Jenna, “How the Machine ‘Thinks:’ Understanding Opacity in Machine Learning Algorithms” (2016) Big Data & Society 1.

Bzdok, Danilo, Naomi Altman & Martin Krzywinski, “Statistics versus Machine Learning” (2018) 15:4 Nature Methods 233.

Cabra Tiago, Sérgio, “Forgetful Al: Al and the Right to Erasure under the GDPR” (2020) 3 European Data Protection Law Review 378.

Calo, M Ryan, “The Boundaries of Privacy Harm” (2011) 86 Indiana Law Journal 1131.

Casey, Bryan, Ashkon Farhangi & Roland Vogl, “Rethinking Explainable Machines: The GDPR's ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise” (2019) 34 Berkeley Technology Law Journal 143.

Castelvecchi, Davide, “Can we Open the Black Box of AI?” (2016) 538 Nature 21.

Chalmers, David, “Facing Up to the Problem of Consciousness” (1995) 2:3 Journal of Consciousness Studies 200.

Challen, Robert et al, “Artificial intelligence, bias and clinical safety” (2019) 28 BMJ Quality & Safety 231.

Chen, Mei & Michel Decary, “Artificial Intelligence in Healthcare: An Essential Guide for Health Leaders” (2020) 33:1 Healthcare Management Forum 10.

Ching, Travers et al, “Opportunities and obstacles for deep learning in biology and medicine” (2018) 15 Journal of the Royal Society Interface 1.

Chui, Michael, James Manyika & Mehdi Miremadi, “Where machines could replace humans—and where they can’t (yet)” (2016) McKinsey Quarterly.

Chynoweth, Paul, "Legal Research" in Andrew Knight & Leslie Ruddock, eds, *Advanced research methods in the built environment* (Chichester: Wiley-Blackwell, 2008).

Collobert, Ronan et al, "Natural Language Processing (Almost) from Scratch" (2011) 12 Journal of Machine Learning Research 2493.

Crépeau, Paul-André, *L'intensité de l'obligation juridique, ou, Des obligations de diligence, de résultat et de garantie* (Montreal: Éditions Yvons Blais, 1989).

Da Silva, Michael, *AI in Health Care: A Fusion of Law & Science* (CIFAR: Toronto, 2021).

De Mauro, Andrea, Marco Greco & Michele Grimaldi, "A Formal Definition of Big Data Based on its Essential Features" (2016) 65:3 Library Review 122.

Deeks, Ashley, "Rulemaking and Inscrutable Automated Decision Tools" 119 Columbia Law Review 1851.

Desai, Deven R & Joshua A Kroll, "Trust but Verify: A Guide to Algorithms and the Law" (2017) 31 Harvard Journal of Law & Technology 1.

Dilhac, Marc-Antoine, Christophe Abrassart, Nathalie Voarino et al, "Montreal Declaration for a Responsible Development of Artificial Intelligence" (Montreal: Université de Montréal, 2018).

Dickens, Bernard, "Medical Negligence" in Jocelyn Grant Downie, Timothy A Caulfield & Colleen M Flood, eds, *Canadian Health Law and Policy*, 4th Edition (Markham: LexisNexis Canada, 2011) 113–151.

Dobrev, Dimiter, "Formal Definition of Artificial Intelligence" (2005) 12 International Journal of Information Theories & Applications 277.

Durán, Juan Manuel & Karin Rolanda Jongsma, "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI" (2020) 47 Journal of Medical Ethics 329.

Edwards Lilian & Michael Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You are Looking For" (2017) 16:1 Duke Law & Technology Review 18.

Eidelson, Benjamin, "Reasoned Explanation and Political Accountability in the Roberts Court" (2021) 130 Yale Law Journal 1748.

Enarsson, Therese, Lena Enqvist & Markus Naarttijärvi, "Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts" (2021) Information & Communications Technology Law 1.

Fiset-Laniel, Julie, Ak'ingabe Guyon, Robert Perreault & Erin C Strumpf, "Public health investments: neglect or wilful omission? Historical trends in Quebec and implications for Canada" (2020) 111 Canadian Journal of Public Health 383.

Fjelland, Ragnar, "Why general artificial intelligence will not be realized" (2020) 7:10 Humanities & Social Sciences Communications 1.

Flood, Colleen & Brian Thomas, "Canadian Medical Malpractice Law in 2011: Missing the Mark on Patient Safety" (2011) 86:3 Chicago–Kent L Rev 1053.

Foss-Solbrekk, Katarina, "Three Routes to Protecting AI Systems and Their Algorithms Under IP Law: The Good, the Bad and the Ugly" (2021) 16:3 Journal of Intellectual Property Law & Practice 247.

Frank, Xavier, "Is Watson for Oncology *per se* Unreasonably Dangerous? Making A Case for How to Prove Products Liability Based on a Flawed Artificial Intelligence Design" (2019) 45:23 American Journal of Law and Medicine 273.

Froomkin, A Michael, 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.

Fuentes, Sigfredo & Eden Jane Tongson, "Implementation of Sensors and Artificial Intelligence for Environmental Hazards Assessment in Urban, Agriculture and Forestry Systems" (2021) 21:6383 Sensors 1.

Gandomi, Amir & Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics" (2015) 35:2 International Journal of Information Management 137.

Geisen, Dieter, *International Medical Malpractice Law: A Comparative Law Study of Civil Liability Arising from Medical Care* (Tübingen: JCB Mohr, 1988).

Gellert, Raphaël, Marvin van Bekkum & Frederik Zuiderveen Borgesius, "The Ola & Uber judgments: for the first time a court recognises a GDPR right to an explanation for algorithmic decision-making" (2021) EU Law Analysis, online: <http://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html>.

Ghassemi, Marzyeh, Luke Oakden-Rayner & Andrew L Beam, "The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care" (2021) 3 Lancet Digital Health 745.

Gibson, Elaine, "Is It Time to Adopt a No-Fault Scheme to Compensate Injured Patients?" (2016) 47:2 OLR 303.

Goodman, Bryce & Seth Flaxman, "European Union Regulations on Algorithmic Decision Making and a 'Right to Explanation'" (2017) 38:3 AI Magazine 50.

Hayit Greenspan et al, "Deep learning in medical imaging: overview and future promise of an exciting new technique" (2016) 35 IEEE Transactions on Medical Imaging 1153.

Grosan, Crina & Ajith Abraham, *Intelligent Systems: A Modern Approach* (New York: Springer Publishing, 2011).

Guidotti, Riccardo et al, "A Survey of Methods for Explaining Black Box Models" (2018) 51:5 ACM Computing Surveys 93:1.

Guly, Christopher, “Data-protection bill needs to be on parliamentary agenda this session, says privacy-law expert,” *the Hill Times* (27 October 2021), online: <https://www.hilltimes.com/2021/10/27/data-protection-bill-needs-to-be-on-parliamentary-agenda-this-session-says-privacy-law-expert/324718>.

Hagen, Gregory, “AI and Patents and Trade Secrets” in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 41–67.

Hagras, Hani, “Toward Human-Understandable, Explainable AI” (2021) 51:9 Computer 28.

Handa, Rish, “Bill 64 Enacted: Québec’s Modern Privacy Regime” McMillan Business Law and Regulatory Law Bulletin (October 15, 2021), online: <https://mcmillan.ca/insights/bill-64-enacted-quebecs-modern-privacy-regime/>.

Haney, Brian S, “The Perils and Promises of Artificial General Intelligence” (2018) 45:7 Journal of Legislation 151.

Hardcastle, Lorian, “Medical Negligence Law in Canada” in Joanna Erdman. Vanessa Gruben & Erin Nelson, eds, *Canadian Health Law & Policy*, 5th Edition (Toronto: LexisNexis, 2017) 305.

Hardcastle, Lorian & Colleen M Flood, “The Future of Health Law: A View Forward from 2016” (2016) Ottawa LR 299.

Harmon, Shawn HE, David E Faour & Noni E MacDonald, “Physician Dismissal of Vaccine Refusers: A Legal and Ethical Analysis” (2020) 13:2 McGill JL & Health 255.

Hassabis, Demis, “Artificial Intelligence: Chess match of the century” (2017) 544 Nature 413.

Haupt, Claudia E, “Governing AI’s Professional Advice” (2019) 64:4 McGill Law Journal 665.

Hemmadi, Murad, “Champagne promises updated privacy legislation in new year” *Regina Leader-Post* (December 6, 2021).

Henglin, Mir et al, “Machine Learning Approaches in Cardiovascular Imaging” (2017) 10:10 Circulation: Cardiovascular Imaging 1.

Hernandez, Karen Andrea Lara et al, “Deep Learning in Spatiotemporal Cardiac Imaging: A Review of Methodologies and Clinical Usability” (2021) 130 Computers in Biology & Medicine 1.

Hilger, Kirsten, Makoto Fukushima, Olaf Sporns & Christian J Fiebach, “Temporal stability of functional brain modules associated with human intelligence” (2019) 41:2 Human Brain Mapping 362.

Huscraft, Grant, “From Natural Justice to Fairness: Thresholds, Content, and the Role of Judicial Review” in Colleen M Flood & Lorne Sossin, eds, *Administrative Law in Context*, 2d ed (Toronto: Emond, 2013) 147–184.

Hutchinson, Terry, “The Doctrinal Method: Incorporating Interdisciplinary Methods in Reforming the Law” (2015) 3 Erasmus Law Review 130.

Jarrahi, Mohammad Hossein, “Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making” (2018) 61:4 Business Horizons 577.
Jordan, MI & TM Mitchell, “Machine Learning: Trends, Perspectives, and Prospects” (2015) 349:6245 Science 255.

Kaminski, Margot E, “The Right to Explanation, Explained” (2019) 34 Berkeley Technology Law Journal 189.

Keane, Pearse A & Eric J Topol, “With an eye to AI and autonomous diagnosis” (2018) 1:40 NPJ Digital Medicine 1.

Kenni, Blen E et al, “Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles” (2019) 7 Institute of Electrical and Electronics Engineers Access 17001.

Kenyon, Miles, “Bill C–11 Explained” (Toronto: Munk School, 2021), online: <https://citizenlab.ca/2021/04/bill-c-11-explained/>.

Khoury, Lara, *Uncertain Causation in Medical Liability* (London: Bloomsbury Publishing, 2006).

Kienle, Gunver S & Helmut Kiene, “Clinical judgement and the medical profession” (2011) 17 Journal of Evaluation in Clinical Practice 621.

Killock, David, “AI outperforms radiologists in mammographic screening” (2020) 17 Nature Reviews Clinical Oncology 134.

Kirkpatrick, Peter, “New clues in the acetaminophen mystery” (2005) 4 Nature Reviews Drug Discovery 883.

Kletzer, Lori G, “The Question with AI Isn’t Whether We’ll Lose Our Jobs — It’s How Much We’ll Get Paid” (2018) Harvard Business Review.

Kononenko, Igor, “Machine Learning for Medical Diagnosis: History, State of the Art and Perspective” (2001) 23:1 Artificial Intelligence in Medicine 89.

Korukonda, Appa Rao, “Taking stock of the Turing test: a review, analysis, and appraisal of issues surrounding thinking machines” (2003) 58 International Journal of Human-Computer Studies 240.

L’Heureux-Dubé, Claire, “The Length and Plurality of Supreme Court of Canada Decisions” (1990) 28 Alberta Law Review 581.

Lacroix, Marième, “Le fait générateur de responsabilité civile extracontractuelle personnelle : continuum de l’illicéité à la faute simple, au regard de l’article 1457 C.c.Q.” (2012) 46:1 Revue juridique Thémis 25.

Lang, Michael, Alexander Bernier & Bartha Maria Knoppers, "AI in Cardiovascular Imaging: 'Unexplainable' Legal and Ethical Challenges?" (2021) Canadian Journal of Cardiology [in press, doi.org/10.1016/j.cjca.2021.10.009].

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton, "Deep Learning" (2015) 521 Nature 436.

Levy, DNL, "How Will Chess Programs Beat Kasparov?" in TA Marsland & J Schaeffer, *Computers, Chess, and Cognition* (New York: Springer, 1990) 47–52.

Liberal Party of Canada, *Forward. For Everyone* (Ottawa: Liberal Party of Canada, 2021), online: <https://liberal.ca/wp-content/uploads/sites/292/2021/09/Platform-Forward-For-Everyone.pdf>.

Lin, Tom CW, "Artificial Intelligence, Finance, and the Law" (2019) 88 Fordham Law Review 531.

Mackie, Tom, "Proving Liability for Highly and Fully Automated Vehicle Accidents in Australia" (2018) 34:6 Computer Law & Security Review 1314.

Malgieri, Gianclaudio & Giovanni Comandé, "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation" (2017) 7:4 International Data Privacy Law 243.

Maras, Marie-Helen & Alex Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos" (2019) 23:3 International Journal of Evidence & Proof 255.

McCarthy, John, "What is Artificial Intelligence" (2004) Stanford University 2, online: <https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf>.

McPhail, Brenda, "Bill C-11 was the gift that needed returning" *the Hill Times* (4 October 2021), online: <https://www.hilltimes.com/2021/10/27/bill-c-11-was-the-gift-that-needed-returning/324436>.

Mécs, André T, "Medical Liability and the Burden of Proof" (1970) 1:16 MLJ 163.

Mendoza, Isak & Lee A Bygrave, "The Right Not to be Subject to Automated Decisions Based on Profiling" in Tatiana-Eleni Synodinou et al, eds, *EU Internet Law: Regulation and Enforcement* (Cham: Springer, 2017) 77–97.

Minsky, Marvin, "Steps Toward Artificial Intelligence" 49:9 (1961) Proceedings of the Institute of Radio Engineers 8.

Moor, James H, "The Status and Future of the Turing Test" 11 (2001) Minds & Machines 77.

Nolan, Donal, "Deconstructing the Duty of Care" (2013) 129 LQ Review 559.

Nosheen, Habiba & Andrew Culbert, "As Fewer Patients Sue their Doctor, the Rate of Winning Malpractice Suits is Dropping too" (2019) CBC News, online: <https://www.cbc.ca/news/health/medical-malpractice-doctors-lawsuits-canada-1.4913960>.

O'Reagan, DP, "Putting Machine Learning into Motion: Applications in Cardiovascular Imaging" (2020) 75 Clinical Radiology 33.

Ordish, Johan & Alison Hall, *Black Box Medicine and Transparency: Interpretable Machine Learning* (Cambridge: PHG Foundation, 2020).

Ozonoff, David, "Legal Causation and Responsibility for Causing Harm" (2005) 95:1 American Journal of Public Health 35.

Parikh, Ravi B, Stephanie Teeple & Amol S Navathe, "Addressing Bias in Artificial Intelligence in Health Care" (2019) 322:24 JAMA 2377.

Pianykh, Oleg S et al, "Continuous Learning AI in Radiology: Implementation Principles and Early Applications" (2020) 297:1 Radiology 6.

Philips-Nootens, Suzanne, Robert P Kouri & Pauline Lesage-Jarjoura, *Éléments de reponsabilité civile médicale: le droit dans le quotidien de la médecine*, 4th edition (Montreal: Éditions Yvons Blais, 2016).

Philp, Mark, "Delimiting Democratic Accountability" (2009) 57 Political Studies 28.

Pinto, Adriano et al, "Combining Unsupervised and Supervised Learning for Predicting the Final Stroke Lesion" (2021) 69 Medical Image Analysis 1.

Popovici, Alexandra, "Le bon père de famille" in Générosa Bras Miranda et Benoit Moore, eds, *Mélanges Adrian Popovici. Les couleurs du droit* (Montréal, Éditions Thémis, 2010) 125–141.

Price III, W Nicholson, "Regulating Black Box Medicine" (2017) 116 Michigan Law Review 421.

Rai, Arun, "Explainable AI: from black box to glass box" (2020) 48 Journal of the Academy of Marketing Science 137.

Raso, Jennifer, "Administrative Law" in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 181–202.

Reddy, Sandeep, Sonia Allan, Simon Coghlan & Paul Cooper, "A governance model for the application of AI in health care" (2020) 27:3 Journal of the American Medical Informatics Association 491.

Renkema, Erik, Manda Broekhuis & Kees Ahaus, "Conditions that influence the impact of malpractice litigation risk on physicians' behavior regarding patient safety" (2014) 14:38 BMC Health Services Research 1.

Rizer, Arthur & Caleb Watney, "Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just" (2018) 23 Texas Review of Law & Politics 181.

Robertson, Gerald B & Ellen I Picard, *Legal Liability of Doctors and Hospitals in Canada*, 5th Edition (Toronto: Thomson Carswell, 2017).

Rollwage, Max et al, “Confidence drives a neural confirmation bias” (2020) 11:2634 Nature Communications 1.

Rothstein, Mark A, “Reconsidering the duty to warn genetically at-risk relative” (2018) 20 Genetics in Medicine 285.

Rudin, Cynthia & Joanna Radin, “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson from An Explainable AI Competition” (2019) 1:2 Harvard Data Science Review 1.

Rudin, Cynthia, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead” (2019) 1 Nature Machine Intelligence 206.

Russak, Adam J et al, “Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype” (2020) 25:5 Journal of Cardiovascular Pharmacology & Therapeutics 379.

Russell, Stuart & Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed (Hoboken: Pearson Education, 2003).

Saygin, Ayse Pinar, Ilyas Cicekli & Varol Akman, “Turing Test: 50 Years Later” (2000) 10 Minds & Machines 463.

Scassa, Teresa, “The Gutting of Consent in Bill C-11” (21 December 2021), online: <http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=336:the-gutting-of-consent-in-bill-c-11&Itemid=80>.

Schifrin, Barry S & Wayne R Cohen, “The Effect of Malpractice Claims on the Use of Caesarean Section” (2013) 27 Best Practice & Research Clinical Obstetrics & Gynaecology 269.

Selbst, Andrew D & Solon Barocas, “The Intuitive Appeal of Explainable Machines” (2018) 87 Fordham Law Review 1085.

Sermesant, Maxime et al, “Applications of artificial intelligence in cardiovascular imaging” (2021) 18 Nature Reviews Cardiology 600.

Shah, Neil et al, “Research Trends on the Usage of Machine Learning and Artificial Intelligence in Advertising” (2020) 19:5 Augmented Human Research 18.

Siegersma, KR et al, “Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist” (2019) 27 Netherlands Heart Journal 403.

Smith, Michael J, “Getting value from artificial intelligence in agriculture” (2020) 60 Animal Production Science 46.

Solove, Daniel J & Danielle Keats Citron, “Privacy Harms” (2021) GWU Legal Studies Research Paper No. 2021-11.

Staszewski Glen, “Reason-Giving and Accountability” (2009) 93 Minnesota Law Review 1253.

Stern, Simon, “Artificial Intelligence, Technology, and the Law” (2019) 68 University of Toronto Law Journal 1.

Stilgoe, Jack, “Machine learning, social learning and the governance of self-driving cars” (2018) 48:1 Social Studies of Science 25.

Sullivan, Hannah R & Scott J Schweikart, “Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?” (2019) 21:2 American Medical Association Journal of Ethics 160.

Swanson, Greg, “Non-Autonomous Artificial Intelligence Programs and Products Liability: How New AI Products Challenge Existing Liability Models and Pose New Financial Burdens” (2019) 42 Seattle University Law Review 1201.

Taddeo, Mariarosaria & Luciano Floridi, “How AI can be a force for good” (2018) 361:6404 Science 751.

Tang, An et al, “Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology” (2018) 69 Canadian Association of Radiologists Journal 120.

Therrien, Daniel, “Submission of the Office of the Privacy Commissioner of Canada on Bill C-11, the *Digital Charter Implementation Act*, 2020” (11 May 2021), online: <https://www.priv.gc.ca/en/opc-actions-and-decisions/submissions-to-consultations/sub_ethi_c11_2105/#toc2-1>.

Therrien, Daniel, “Building back better requires strong, effective regulation of digital world,” *the Hill Times* (27 September 2021), online: <https://www.hilltimes.com/2021/09/27/building-back-better-requires-strong-effective-regulation-of-digital-world/319480>.

Thomasen, Kristen, “AI and Tort Law” in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 103–122.

Thorogood, Adrian, Alexander Bernier, Ma’n H Zawati & Bartha Maria Knoppers, “A Legal Duty of Genetic Recontact in Canada” (2019) 40:2 Health Law in Canada 58.

Tian, Lingling, Juncheng Jiang & L Tian, “Safety analysis of traffic flow characteristics of highway tunnel based on artificial intelligence flow net algorithm” (2019) 22 Cluster Computing 573.

Tom, Elyse et al, “Protecting Data Privacy in the Age of AI-Enabled Ophthalmology” (2020) 9:2 Translational Vision Science & Technology 1.

Topol, Eric, “High-performance medicine: the convergence of human and artificial intelligence” (2019) 25 Nature Medicine 44.

Trister, Andrew Daniel, “The Tipping Point for Deep Learning in Oncology” (2019) 5:10 JAMA Oncology 1429.

Turing, AM, “Computing Machinery and Intelligence” (1950) 59:236 Mind 433.

Vamplew, Peter et al, “Human-aligned artificial intelligence is a multiobjective problem” (2018) 20 Ethics and Information Technology 28.

Vollmann, Jochen & Rolf Winau, “Informed consent in human experimentation before the Nuremberg code” (1996) 313 British Medical Journal 1445.

Voosen, Paul, “How AI detectives Are Cracking Open the Black Box of Deep Learning” (2017) Science Newsletter, online: <https://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning>.

Wachter, Sandra, Brent Mittelstadt & Chris Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR” (2018) 31:2 Harvard Journal of Law & Technology 841.

Wachter, Sandra, Brent Mittelstadt & Chris Russell, “Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI” (2021) 41 Computer Law & Security Review 1.

Wachter, Sandra, Brent Mittelstadt & Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation” (2017) 7:2 International Data Privacy Law 76.

Wang, Fei, Lawrence Peter Casalino & Dhruv Khullar, “Deep Learning in Medicine: Promise, Progress, and Challenges” (2018) 179:3 Health Care Reform 293.

Weinrib, Ernest J, *Tort Law: Cases & Materials*, 5th Edition (Toronto: Emond Montgomery Publications, 2019).

Weinrib, Ernest J, *Corrective Justice* (Oxford: Oxford University Press, 2012).

Weinrib, Ernest, “A Step Forward in Factual Causation” (1975) 38 Modern Law Review 518.

Zador, Anthony M, “A Critique of Pure Learning and what Artificial Neural Networks Can Learn from Animal Brains” (2019) 10 Nature Communications 1.

Zednik, Carlos, “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence” (2021) 34 Philosophy & Technology 265.

Jurisprudence

Amsterdam District Court, Amsterdam, 11 March 2021, *Applicant v Ola Netherlands BV*, C/13/689705, HA RK 20-258.

Baker v Canada (Minister of Citizenship and Immigration), [1999] 2 SCR 817, 174 DLR (4th) 193.

Barnett v Chelsea and Kensington Hospital Management Committee, [1968] 1 All ER 1068, [1969] 1 QB 428.

Benhaim v St. Germain, 2016 SCC 48, [2016] 2 SCR 352.

British Columbia v Canadian Forest Products Ltd, 2018 BCCA 124, CA43841.

Clements v Clements, 2012 SCC 32, [2012] 2 SCR 181.

Cooper v Hobart, 2001 SCC 79, [2001] 3 SCR 537.

Crits v Sylvester, [1956] OR 132 at 13, 1 DLR (2d) 502 (Ont. CA).

Cuthbertson v. Rasouli, 2013 SCC 53, [2013] 3 SCR 341.

Donoghue v Stevenson, [1932] UKHL 100, [1932] All ER Rep 1.

Halushka v University of Saskatchewan, [1965] 53 DLR 2nd 436.

HH v RG, Health Professionals Appeal & Review Board, ON (2013), 11–CRV–0178.

Hopp v Lepp, [1980] 2 SCR 192, 13 CCLT 66.

Lapointe c Hôpital Le Gardeur, [1992] 1 SCR 351 at 363, 90 DLR (4th) 7.

Mustapha v Culligan of Canada Ltd, 2008 SCC 27, 293 DLR (4th) 29.

Overseas Tankship Ltd (UK) v Miller Steamship Co Pty, [1967] AC 617 (PC), [1967] 2 All ER 709.

Reibl v Hughes, [1980] 2 SCR 880, 114 DLR (3d) 1.

Salgo v Leland Stanford, (1957) 154 Cal App 2d 560, 317 P2d 170.

Snell v Farrell, [1990] 2 SCR 311 at 320, 72 DLR (4th) 289.

St–Jean v Mercier, 2002 SCC 15, [2002] 1 SCR 491.

Starson v Swaze, 2003 SCC 32, [2003] 1 SCR 722.

United States v Karl Brandt et al (Doctor's Trial, 1947), in *Records of the United States: Nuremberg War Crimes Trial* (Washington: National Archives and Record Service, 1974).

Watters v White, 2012 QCCA 257, JE 2012-473.

Wilson v Swanson, [1956] SCR 804 at 812, 5 DLR (2d) 113.

X v Mellen, [1957] BR 389.

Legislation & Normative Documents

An Act to modernize legislative provisions as regards the protection of personal information, CQLR, c 25.

"An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts," *House of Commons Debates*, 43-2, 150, no 30 (17 November 2020).

"An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts," *House of Commons Debates*, 43-2, 150, no 35 (24 November 2020).

Article 29 Data Protection Working Party, "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679" (EU) 2017.

Bill C-11, *An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts*, 2nd Sess, 43rd Parl, 2020.

Canadian Medical Association, *Code of Ethics and Professionalism*, Ottawa: CMA, 2018.

Charter of the French Language, CQLR, c 11.

Code de déontologie des médecins, CQLR, M–9, C–26, 2002 r 17.

Code de la santé publique, JO, 2 août 2021 (NC).

EC, *General Data Protection Regulation* (EU) 2016/679, OJ L 119.

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*, OJ, COM(2021) 206 final, 2021/0106(COD).

Health Canada, "Guidance Document: Software as a Medical Device (SaMD): Definition and Classification" (Ottawa: Her Majesty the Queen in Right of Canada, as Represented by the Minister of Health, 2019).

Health Care Consent Act, RSO 1996, c 2, Sched A.

Loi no 2018-493 du 20 juin 2018 relative à la protection des données personnelles, JO, 21 June 2018, JUSC1732261L.

Loi modernisant des dispositions législatives en matière de protection des renseignements personnels, CQLR, c 25.

OECD, Council on Artificial Intelligence, *Recommendation of the Council on Artificial Intelligence* C(2019)34, C/MIN(2019)3/FINAL.

Ontario, White Paper, *Modernizing Privacy in Ontario Empowering Ontarians and Enabling the Digital Economy* (Toronto: Government of Ontario, 2021).

Quebec, National Assembly, *Journal des débats (Hansard)*, 42-1, vol 45, no 120 (12 juin 2020).

Quebec, National Assembly, *Journal des débats of the Committee on Institutions (Hansard)*, "Clause-by-clause consideration of Bill 64, An Act to modernize legislative provisions as regards the protection of personal information" 42-1, vol 45, no 113 (2 février 2021).

Treasury Board Secretariat, "Directive on Automated Decision-Making" (Ottawa: Her Majesty the Queen in Right of Canada, represented by the President of the Treasury Board, 2019).
UNESCO, *First Draft of the Recommendation on the Ethics of Artificial Intelligence*, SHS/BIO/AHEG-AI/2020/4 REV.2, Paris (2020).

Commercial Online Resources

Arterys, “About Us” (2021), online: <https://arterys.com/about-us>.

Canon Medical Systems, “Advanced intelligent Clear-IQ Engine (AiCE)” (2020), online: <https://global.medical.canon/products/magnetic-resonance/aice>.

GE Healthcare, “Critical Care Suite 2.0” (2021), online: <https://www.gehealthcare.com/products/radiography/mobile-xray-systems/critical-care-suite-on-optima-xr240amx>.

Hsieh, Jiang et al, “A new era of image reconstruction: TrueFidelity™ (Technical White Paper, JB68676XX, 2019).

iCAD, “Artificial Intelligence for Digital Breast Tomosynthesis: Reader Study Results” (White Paper, DMM253 Rev B, 2020).

Icometrix, “Enabling value-based care for people with neurological conditions” (2021), online: <https://icometrix.com>.

Medtronic, “Guardian™ Connect Continuous Glucose Monitoring (Cgm) System Now Licensed in Canada for People Living with Diabetes” (May 23, 2018), online: <https://www.medtronic.com/ca-en/about/news/Guardian_Connect_Press_Release.html>.

Siemens Healthineers, “AI-Rad Companion” (2021), online: <https://www.siemens-healthineers.com/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion>.