Ethical Theory and Moral Practice
https://doi.org/10.1007/s10677-023-10390-4

Explainability, Public Reason, and Medical Artificial Intelligence

Michael Da Silva[1,2]

Accepted: 30 April 2023

Abstract
The contention that medical artificial intelligence (AI) should be 'explainable' is widespread in contemporary philosophy and in legal and best practice documents. Yet critics argue that 'explainability' is not a stable concept; non-explainable AI is often more accurate; mechanisms intended to improve explainability do not improve understanding and introduce new epistemic concerns; and explainability requirements are ad hoc where human medical decision-making is often opaque. A recent 'political response' to these issues contends that AI used in high-stakes scenarios, including medical AI, must be explainable to meet basic standards of legitimacy: People are owed reasons for decisions that impact their vital interests, and this requires explainable AI. This article demonstrates why the political response fails. Attending to systemic considerations, as its proponents desire, suggests that the political response is subject to the same criticisms as other arguments for explainable AI and presents new issues. It also suggests that decision-making about non-explainable medical AI can meet public reason standards. The most plausible version of the response amounts to a simple claim that public reason demands reasons why AI is permitted. But that does not actually support explainable AI or respond to criticisms of strong requirements for explainable medical AI.

Keyword Political Philosophy · Artificial Intelligence · AI · Governance · Public Reason

Concerns about 'black box medicine' (Price 2015) undergird criticisms of many medical artificial intelligence (AI) tools.[1] AI has tremendous potential in healthcare settings, potentially promising more accurate and efficient decisions, shorter wait times, and more efficient resource allocations (Topol 2019a, b). Yet many promising tools are

[1] AI here "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" (OECD 2019: 4). ML can "adapt to new circumstances … to detect and extrapolate patterns" (Russell and Norvig 2021). Deep learning is a species of ML with many processing nodes, including 'hidden' ones. See also Bringsjord and Govindarajulu (2018).

✉ Michael Da Silva
M.da-silva@soton.ac.uk

[1] School of Law, University of Southampton, Southampton, UK

[2] Senior Fellow in AI and Health Care, AI + Society Initiative, University of Ottawa, Ottawa, Canada

non-transparent. Consider, e.g., deep learning-enabled tools aimed at early suicidal ideation detection (e.g., Roy et al. 2020) or bespoke recommendations for treatment-resistant depression (e.g., Pigoni et al. 2019). These could provide more accurate, efficient mental healthcare absent understanding of precise mechanisms by which they do so – and perhaps even absent high probabilities about which inputs contributed most to an outcome and a plausible story of how.[2] Yet many understandably worry that the opacity of many AI (especially deep learning)-enabled tools can, e.g., lead to missed AI-induced errors and biases or undermine provision of justificatory reasons for medical decisions and, consequently, trust in AI. Some champion explainable AI (XAI) (defined below).

Clinicians, patients, and innovators claim "rights" to understand medical AI (Panch et al. 2019). Ursin et al. (2022) suggest all bioethical principles require explainability. XAI advocacy informs developments in academia, law/policy, and industry. Many academics (e.g., Lundberg et al. 2020; Yap et al. 2021) promote XAI. Some (e.g., Maclure 2021) even propose "strong" explainability requirements under which only AI above an explicability threshold should be available. Many policy-makers accept these arguments. To wit, the *EU General Data Protection Regulation* includes a legal right to explanation and explainability is part of Ontario (2022)'s first principle for ethical AI. XAI development thus unsurprisingly continues (Arrieta et al. 2020).

This work nonetheless defends the use of at least lower risk non-explainable health-related AI subject to rigorous performance testing and value of administrative decision bodies for evaluating non-explainable medical AI. It specifically defends such use against a new political critique, most forcefully articulated by Maclure (2021), suggesting non-explainable AI cannot meet Rawlsian public reason standards (Rawls 1993).[3] I demonstrate that levels of transparency and accountability required by public reason in medical settings do not support strong XAI requirements and an administrative state can provide checks on medical AI accuracy and safety that meet proper political standards. My arguments aim to demonstrate that AI need not be explainable in any strong sense to be permissibly provided on markets or used by medical professionals. I do not claim that explainability tools have no role in any model-auditing required for proper safety and efficacy review. Rather, I contend that some AI need not conform to strong explainability requirements to pass the kinds of review demanded by public reason standards and these AI can be justifiably distributed, used, and perhaps even funded if they pass such review.

I first outline classic arguments for explainable medical AI. I then survey existing and identify novel issues with calls for explainable medical AI, thereby disentangling prominent arguments. I next outline the political response. I present two readings and show neither succeeds. On one, justice demands decisions based on knowledge of how the AI operates. This raises problems parallel to those with other calls for explainability and novel issues. On another, justice demands provision of reasons for why AI is permitted in a sphere. This straightforwardly applies existing norms from outside AI, does not speak to AI opacity, and thus does not respond to positions it ostensibly negates. I finally address objections in a broader argument for non-explainable AI.

This analysis has practical and theoretical implications for AI, healthcare, and governance. It aims to push "forward not only the academic debate in computer science and other disciplines, but also that within philosophy" by further illuminating

[2] If one dislikes these examples– and I note limits in Pigoni et al. (2019) –see many others in overviews of medical AI (e.g., Topol 2019a, b) and analyses thereof (Price 2018; London 2019).

[3] Aspects also appear in works demanding explainability as a matter of justice. Further to texts above, see e.g., Amann et al. (2020). Vredenburgh (2022) provides a non-medical, rights-based political response.

"basic issues and concepts" (Zimmermann et al. 2022: 2). Practically, it contributes to debates regarding whether non-explainable AI should be permitted on markets and whether there should be legal rights to explanation (Watson and Floridi 2021: 9213). The underlying logic impacts the scope of such a right: it could support a form of healthcare exceptionalism treating medical technologies differently than other AI-enabled products or demonstrate that not all medical AI should be considered high-stakes if that triggers explainability requirements. Theoretically, the analysis implicates issues concerning whether there is a moral right to explanation and how it implicates stakeholders' duties. Arguments also clarify questions of legitimacy in administrative states and the meaning of public reason in liberal-democratic administrative states tasked with addressing global technological changes. Even if my arguments fail, data therein should be useful for inquiries into whether and how we should treat parallels between human and AI 'decision-making.'

1 Background/Terminology

To begin with definitions, AI is 'explainable' here where one can understand how and why it operates to reach decisions. XAI is generally taken to address concerns with AI 'opacity' defined as AI's being "difficult to understand why it does what it does or to know how it works" (Zednik and Boelsen 2022: 220).[4] AI can be opaque for several reasons. Some problems only apply to some AI. Consider how several layers of deep learning algorithms can hide relevant processing nodes from view or how 'adaptive' machine learning (ML) tools can have unpredictable changes in how they preform and operate over time in response to new data (Price 2015, 2017, 2018). Other problems apply generally. Consider concerns that knowledge is lost as terms are translated from natural language to code and back again, obscuring decision-making (Pierce et al. 2022). XAI advocates nonetheless share desires for circumstances where stakeholders can understand how and why AI provided with a dataset A at point X made decision B.

This broad view of XAI and its aims captures most phenomena in arguments for explainable medical AI. However, it admits distinctions, including those between systems-level (how a tool operates generally) and individual-level (how it works in particular cases) and model-specific and model-generic explanations (Watson 2021). One concerns 'interpretable' AI and XAI. The terms are sometimes used interchangeably and authors distinguish them differently (Watson and Floridi 2021: 9212; Watson 2022a). For example, Babic et al. (2021) suggest interpretable AI uses transparent functions. Its inputs, weights, etc. are each understandable. So, we can understand how decisions are reached. XAI, by contrast, uses one AI model to get indirect 'understanding' of another, opaque one. For instance, one can create a similar, transparent algorithm that fits outputs, rather than original training data. Watson and Floridi (2021) suggest focusing on interpretable AI to emphasize "the subjective goal of interpretation over the (ostensibly) objective goal of explanation" and distinct ML-related issues. Yet Zerilli (2022) defines interpretability in terms of agent-level/

[4] The latter concern is sometimes framed as pertaining to a lack of information about whether and how its inputs reach decisions– or "link uncertainty" whereby the connection between purported support for a decision and the decision itself is unclear (Sullivan 2022) –severable from opacity concerns strictly defined. Notably, link uncertainty often stems from issues outside the model itself. I will return to some issues this presents below.

folk-psychological understanding and suggests it is just one of XAI's three aims.[5] Herzog (2022) then notes that 'intelligibility' is used as a "generic" encompassing 'explainability' and 'interpretability' but suggests "explicability" is a better generic focused on "explanations that can be understood and utilized in practice."[6]

XAI proponents offer different epistemic targets (interpretability, transparency, understanding, trustworthiness, etc.) acceptable medical AI must meet (Arrieta et al. 2020). The issue they seek to address goes by several names, including the "black box," "explainability," "transparency," "interpretability," or "intelligibility" problem(s) (Maclure 2021: 422). Some seek to explain "what" an AI tool does while others seek to explain "why" it produces a result (Zednik 2021: 274). Different stakeholders– AI developers, users, regulators, etc. –may need to be in different epistemic states to fulfill their different purposes (Tomsett et al. 2018; Zednik 2021). The types of explanation appropriate to those tasks may further differ. One may require distinguish model-centered explanations used for purposes like auditing, clinician-centered explanations that could support decision-making, and patient-centered ones focused on trust and informed consent.[7] Yet XAI advocates each seek explanations that "tell us why x is true" (Watson et al. 2019) where x represents input–output relations. Purported problems arise where it is unclear how AI "arrived at a particular output," undermining trust, acceptance, and responsibility (Ratti and Graves 2022).

This work uses XAI as its generic for argumentative purposes. Little turns on whether 'XAI' best describes the phenomena at issue. Calls for XAI take many forms but share basic commitments to the claim that AI should be available iff one can explain how and why it reaches decisions. More specifically, most consider AI "explainable" iff one can provide an account of mechanisms by which it reached a decision, the rules undergirding that decision-making, and how particular inputs produced a decision when fed into the AI to trigger the rules. At minimum, they contend, XAI admits "high-level" explanations of how and why set inputs produced particular outcomes.[8] This does not require a *complete* causal story of how and why each element of a tool contributed to each outcome. However, if the AI is to qualify as explainable, one should be capable of providing a general account of its operations– through transparent design or use of secondary algorithms –that (at least) identifies the elements most likely to be responsible for an outcome and plausible reasons why the AI would view them as apt. This minimalist account of explainability will remain controversial but captures many strong explainability requirement advocates' self-professed desires while providing them with a wider range of argumentative moves than are available on other accounts. It thus reflects the principle of charity. Consistent with this, I am agnostic here on whether the method of identifying how AI reaches decisions is systems- or individual-level, ex-ante or ex-post, and even the precise technical data required.

XAI so-defined takes many forms. For example, Babic et al. (2021)'s distinction tracks, and can be reformulated as, another between 'inherent' (or 'intrinsic') explainability and 'post-hoc' explainability (Watson 2021: 49ff; Ghassemi et al. 2021). Inherent explainability is ex-ante transparency regarding how AI reaches decisions such that relationships between AI inputs and outputs are clear. Inherently explainable AI do not raise clear intelligibility issues; one can always explain how inputs produce outputs (Watson 2021: 49). They are most likely to work when comprising "simple models with clear internals, leading to widespread attempts to use post-hoc explanations for potentially complex, black-box

[5] The others are completeness/depth and fidelity. Compare, e.g., Watson and Floridi (2021: 9222-9225).

[6] Zerilli (2022) also distinguishes "fathomable" AI systems, whose functions one can understand in full, and intelligible ones, of which one can gain understanding of some parts via inspection.

[7] I thank an anonymous reviewer for this distinction.

[8] Vredenburgh (2022) discusses this in terms of rule-based causal and normative explanations for decisions.

models" (Poursabzi-Sangdeh et al. 2021). Post-hoc explainability tools try to understand how an AI tool 'decided.' They include the use of more transparent secondary AI to reproduce the original (but otherwise-opaque) AI's results or provide statistical probabilities of how much an input contributed to those results.

Many post-hoc explainability tools simplify AI processes into more complex versions of inherently explainable AI tools, like linear regressions or rule-lists. Consider "feature attribution methods," which seek to identify the extent to which inputs produced particular outputs (*id*.). The sophisticated linear regression and rule-lists "attempt to approximate some complex functional relationship with an alternative method considered more readily interpretable" (*id*.: 49). Local interpretable model-agnostic explanations (LIME) and Shapley Additive Explanations (SHAP), for example, provide linear regressions that approximate and best explain otherwise-opaque tools, like deep neural network models, reached decisions (*id*.; Watson 2022b: 1503). Rule-lists intend to serve the same function; they are often visualized as decision trees with a series of "if–then statements" (Watson 2022b: 1506–1507; Zerilli 2022; Zednik 2021: 277). Such tools provide 'explanations' in different ways. SHAP, for instance, is part of a family of post-hoc techniques, including visualizations, statistical analyses, etc., that aim to "identify high-responsibility inputs" in systems (Zednik and Boelsen 2022: 222). Other examples include heatmaps, which assign values for each input's responsibility for an output and represents them visually in a colour-coded 'map' of a decision space, and feature-detector visualization, which represents key inputs pictorially (Zednik 2021: 275–276, 281). LIME, by contrast, is a form of "surrogate modeling" where one seeks understanding via secondary AI that approximates opaque AI (Zednik and Boelsen 2022: 221). Still other explainability tools measure the extent to which input layers impact outcomes, including via comparisons with outcomes in the surrounding environment (as in layer-wise relevance propagation) (*id*.; Zednik 2021).

Some XAI tools do not provide complete causal stories of the input–output relationships one may expect of a stronger mechanistic account of 'how' decisions are reached, but each of the forgoing could help meet the more capacious definition of explainability above. If they do not submit to that definition, they are beyond my scope of inquiry. Whether my phenomenon best fits use of the term 'XAI' is, again, debatable. But this mechanistic account has precedent in multiple domains, including in work by key figures in the present debate like Maclure (2021) and London (2019). At minimum, strong explainability advocates in the medical sphere seek knowledge of this kind to provide professionals with guidance as to whether the AI is likely to be making decisions on proper bases and inform patients about same. This is taken to require some understanding of how and why AI reached its decision. Even purported XAI that do not aim to and cannot provide a complete causal story of how AI operates seek to fulfill this basic task.[9]

[9] As a reviewer rightly notes, SHAP values, for instance, merely provide summaries of key predictors on a game theoretical model of how AI with the same inputs may have reached its results. This provides some indication of how inputs led to particular outcomes but SHAP values alone do not identify the causal mechanism that links them. If SHAP values cannot contribute to that story or advocates for their use do not seek any mechanistic story, calls for 'XAI' of this kind may be beyond the present scope of inquiry. My view could be read as responding to a different set of calls for strong mechanistic explanations. However, SHAP, LIME, and other tools providing indications of the extent to which certain inputs contributed to certain outputs appear relevant to the task at hand. The majority of views promoting strong explainability requirements are likely only interested in AI of these forms where they provide a basic for further reasonable inferences about whether and why these predictors are acceptable. SHAP values (and similar XAI, like LIME) are part of a broader explainability process that builds on their identification of the elements most likely to be responsible for an outcome and seek plausible reasons why it would take these to be apt. They may be valuable parts of such a process. Whether they are required for it remains less clear.

Explainable medical AI advocates, then, seek to ensure (at least) medical professionals can understand mechanisms AI use to reach decisions. Many advocates further desire knowledge of causal relationships in a domain and principles explaining them (London 2019).[10] Methods designed for AI 'explainability' aim to establish robust relationships between inputs and outputs, again suggesting that XAI advocates want to explain how decisions are reached. While advocates may not require a complete causal story, they seek enough information to identify the mechanisms, rules, and inputs most likely to have produced the outcome and assurances those are well-chosen. The question is whether *all* medical AI should be explainable in this sense. I will briefly explain why many believe they should before motivating my negative response.

2 Classic Arguments for Explainable Medical AI in Brief

Explainable medical AI claims numerous benefits. Some contend that explainability requirements improve performance by, e.g., eliminating causal noise or identifying "perturbations" that could undermine robustness (Arrieta et al. 2020: 83) or help identify/fix 'bugs' (Poursabzi-Sangdeh et al. 2021; Watson and Floridi 2021: 9212–9213) or technical issues, like "overfitting" whereby AI is too accurate for one dataset in ways that suggest its solutions cannot generalize (Watson/Floridi *id.*).[11] Others suggest XAI best identifies safety risks, mechanisms that lead to AI-induced harms, and methods of addressing them: it is hard to address problems one cannot predict ex-ante or understand ex-post (*id.*; Arrieta et al. 2020: 83; Poursabzi-Sangdeh et al. 2021; Ghassemi et al. 2021; Yoon et al. 2022). Biases and related data issues may also be easier to detect and address if AI is explainable (Yoon et al. *id.*; Watson and Floridi 2021: 9236).

Further arguments suggest XAI fosters transparency and more broadly helps justify AI-related decisions, which is required where decisions will impact vital interests, or promotes trust in AI, which is necessary to secure its potential benefits (compare e.g., London 2019; Ghassemi et al. 2021; Maclure 2021). People are owed explanations for decisions that impact them and will not accept even beneficial AI use without such decisions (see also Poursabzi-Sangdeh et al. 2021).

Explainable medical AI, then, purportedly furthers many goods related to intrinsic epistemic or political goods, basic safety, and other instrumental goods, like public trust. Calls therefor are unsurprising. Stronger versions champion the exclusive use of explainable medical AI and even legal standards that would bar the use of non-explainable AI tools in healthcare settings.

3 Issues with Explainable Medical AI

Arguments for explainable medical AI nonetheless face issues. I will outline the general issues first before moving on to examine the new political one at issue here. It is, for instance, difficult to identify the relevant epistemic or moral target. Call this 'the argument from imprecision.' Arguments for 'explainable AI,' again, refer to different epistemic standards. If one

[10] Watson (2021: 170) suggests *all* relevant explanations are causal.

[11] They further suggest it can lead to scientific discoveries. See also Zednik and Boelsen (2022); Herzog (2022).

focuses on "understanding" as the basic norm (Arrieta et al. 2020; Sullivan 2022), there is still no good technical definition of what this should require of medical AI (Ghassemi et al. 2021). Legal standards, like the E.U.'s right to "meaningful information about the logic behind automated decisions," do not explain whether we should seek inherent or post-hoc explainability or help specify technical standards required to meet either standard. Whether AI needs to be in principle capable of being understood by a relevant party (e.g., a physician) or actually understood by some population (e.g., most physicians) to properly qualify as 'explainable' also remains unclear.

Concerns that different audiences require different kinds of explanations (Arrieta et al. 2020) also raise challenges. Levels of explanation required for regulators to permit a sale of a tool and for a physician to use it can differ. XAI advocates grant this, but the problem posed is greater than many suggest: Specifying *all* relevant epistemic standards is a tall order where we now lack *any* precise ones. And, as we will see below, the levels of understanding plausibly required for various purposes do not generally require explanations of mechanisms undergirding decisions. This is so for XAI advocates' model-centered, clinician-centered, and patient-centered goals.

The goods explainability should serve are likewise unclear. Are they intrinsic or instrumental? Epistemic or moral?[12] It is hard to know what explainability is supposed to mean or assess arguments therefor absent clarity on what explainability should promote. Instrumental arguments require empirical support proponents of stronger requirements are unlikely to adduce. Explainability is, e.g., neither necessary nor sufficient to address bias concerns or protect privacy. Even those who appeal to bias concerns grant that explainability would only make addressing it *easier* (Yoon et al. 2022). Other means of addressing it could prove technically or otherwise superior. After all, Obermeyer et al. (2019) identified biases in non-explainable AI.

If more sophisticated accounts address the preceding, empirical considerations present further challenges. One stems from ways in which explainability requirements undermine medical AI's potential benefits. Call this 'the argument from accuracy.' Even the best healthcare systems suffer from high rates of iatrogenic injury, long wait times, high medical costs, and provider bias (Da Silva et al. 2022). Advances in AI across the spectrum of care offer prospects of increased accuracy, increased efficiency, and, if properly calibrated, could mitigate provider biases (*id.*). These goods rely on AI performance that does not correlate with explainability enough to justify a strong requirement. There is "no logical or statistical guarantee that interpretable models will outperform black box competitors or even be in the Rashomon set of high-performing models for any given predictive problem" (Watson and Floridi 2021: 9235). The most effective AI are not the most explainable ones (London 2019; Babic et al. 2021). Non-explainable deep learning AI is responsible for many especially exciting recent medical developments, including the mental health examples above.[13]

Inherent explainability requirements give up the possibility of deep learning (Babic et al. 2021). Simple models do not present serious accuracy issues (Poursabzi-Sangdeh et al. 2021). Yet inherently explainable deep learning is unlikely to prove possible. Non-explainable deep learning is, again, already valuable in healthcare settings and has tremendous future potential. Post-hoc measures may not fill remaining knowledge gaps key to effective deployment. Many do not focus on real-world accuracy but on creating

[12] See also Arietta et al. 2020 (listing "Trustworthiness;" "Causality;" "Transferability;" "Informativeness;" "Confidence;" "Fairness;" "Accessibility;" "Interactivity;" and "Privacy Awareness").

[13] See also examples in note 2 sources.

'white box' algorithms that fit the 'black box' AI (Babic et al. 2021). They compound inaccuracy risks: the primary algorithm and white box AI that could explain it both offer opportunities for technical error (Ghassemi et al. 2021). Heatmaps and other non-surrogate modelling tools do not, by contrast, rely on white boxes. Future innovations could, theoretically, minimize the necessity of many existing trade-offs between 'interpretability' and performance (Rudin 2019). However, even the best existing and proposed post-hoc tools raise non-de minimum accuracy trade-off concerns and reliance on secondary algorithms still introduces opportunities for new misunderstanding/errors. And if heatmaps, etc. avoid *these* concerns, advocates still grant that their explanations cannot fulfill many purposes.

The second empirical argument, 'the argument from the limitations of explainability mechanisms,' notes that many forms of XAI do not achieve their epistemic or moral goals. Mixed existing evidence here cannot support strong explainability requirements. For example, a recent study suggests simple models with few features could improve understanding but do not lead persons to follow recommendations more closely *when it would be beneficial* and undermine abilities to identify/correct substantial mistakes (Poursabzi-Sangdeh et al. 2021). Babic et al. (2021) suggest post-hoc explainability tools do not provide sufficiently better understanding of how original devices work and produce false senses of ("ersatz") understanding that maintain opacity-related problems. Ghassemi et al. (2021) add that many explainability tools (heatmaps, contextual language models, etc.) demonstrate *what* was relevant to a decision but not *why* it was relevant or whether it should have been. Even XAI advocates admit it only provides goods under certain conditions. For example, Zednik (2021: 286) suggests heatmaps answer 'what' and 'why' questions and feature-detector visualization can answer 'how' and 'where' questions but both work best where underlying data submit to semantic interpretability. While distinct kinds of explanation could, again, prove appropriate for different stakeholders, many tools' explanations are not like those relevant to clinical decisions (Lindsell et al. 2020). Primary AI may accordingly still fail to meet any standard for use proposed by strong XAI advocates.

More positive data cannot justify strong explainability requirements. For instance, Tschandl et al. (2020) report in a letter that human raters in their study were more accurate when informed by an explainable 34-layer neural network and propose that "explanations for AI-based predictions can be translated into a human-understandable visual concept" (*id*.: 1232). Yet they also report persons working with their explainable tool were "vulnerable to perform below their expected ability if there is a fault with the AI. Whether techniques to facilitate interpretability or explainability mitigate the risk of this negative impact" was unclear. Another study suggests explainability tools can provide an "interface" between humans and AI to, e.g., translate lymph node diagnostic tool results into classifications that can be "integrated in the wider 'medical culture' of diagnosis" (Ratti 2022). Yet explainability tools are not the only possible interface. Even the most effective ones may not work well for everyone. Multiple major post-hoc explainability tools present significant variation across protected groups (Balagopalan et al. 2022), undermining claims that explainability requirements will address biases. While Watson (2022a) optimistically believes future AI will address many issues, challenges remain.

Ghassemi et al. (2021) further note that existing post-hoc explainability tools' inability to identify precisely how systems operate and reach particular decisions lets humans fill in narrative gaps and provides them with cover to make decisions that they would make anyway. This mirrors common concerns that AI generally provides a veneer of neutrality to deeply political decisions that should be openly contested (e.g., Benjamin 2019 (primarily on bias)). If the goal is better reasons for decisions,

post-hoc measures will not provide them and can give a veneer of technologically-informed justification to otherwise problematic claims. If the goal is accountability, at least post-hoc ersatz understandings simply do not permit proper 'checks.' Where, moreover, the meaning of accountability in AI settings is still in development (Johnson 2021), blunt explainability requirements may limit opportunities to develop more nuanced views.

Increased transparency is not clearly necessary or sufficient for the understanding required to full relevant ends. Sullivan (2022) suggests the real problem with many ML tools is "a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon." The relevant epistemic state requires understanding of the relationship between the algorithm and underlying phenomena that it models, not merely how input set X produces outcome Y. Indeed, Sullivan suggests, one can suitably understand 'opaque' AI if "there is an adequate link connecting the model to the phenomenon of interest." If so, 'explainability' requirements could constitute a red herring. One needs a better means of ensuring algorithms actually model real-world phenomena. XAI is not obviously required for this either.

Sullivan's broader concern with a "link uncertainty" concerning misalignment(s) between our understanding of a model's functionality and external evidence remains.[14] However, at least many link uncertainty-related issues could be better addressed by stronger evidentiary standards throughout medical research development. Sullivan, for example, notes that many issues with link uncertainty in medical cases stem from the lack of strong empirical data about various medical conditions to feed into the medical AI (*id.*: 124–125). She thus suggests using deep learning-enabled AI to identify new research questions, rather than directly informing care. Yet higher evidential standards for both underlying data quality and AI performance minimize risks of link uncertainty and any attendant harms if AI is available in any case. There is further reason to raise evidentiary standards for safety and efficacy review in many sectors and to be cautious about AI absent very strong evidence as proper standards continue to develop (Da Silva et al. 2022). However, strong explainability requirements are not obviously necessarily for these purposes. Whether they are necessary to address related empirical issues remains unclear. At best, the above/below suggests only some XAI are likely to solve more problems than they raise.

An 'argument from unintended consequences' then notes that mandating XAI use could backfire. Simkute et al. (2021) find that using algorithmic systems can disrupt domain experts' ability to use their expertise; non-tailored explainability mechanisms can result in automation bias and algorithmic aversion. Individuals could, more broadly, prove more susceptible to automation bias when using XAI, giving it undue deference and so not correcting errors. Poursabzi-Sangdeh et al. (2021) note that persons in their study failed to identify and correct mistakes even on more simple models. Jacobs et al. (2021) suggest provision of *any* explanation of how AI operates or reaches decisions increases chances that healthcare providers will follow inaccurate AI recommendations.[15] Even Tschandl et al. (2020) found that users of their XAI tool began to ignore it over time. While empirics here are contestable, concerns that XAI requirements may undermine safety by providing false senses of understanding are presently reasonable.

[14] I thank an anonymous reviewer for helping me to see this point and some phrasing here.

[15] Following note 2, Jacobs et al. (2021)'s psychiatric case underlines non-explainable AI's value in mental health.

Calls for XAI are, moreover, curious where human decision-making is also opaque. The 'argument from the limitations of human reasoning' (Maclure 2021) notes that AI is no worse at meeting relevant epistemic standards than humans. Human medical decision-making often takes place without clear understandings of the mechanisms by which treatments, etc. work. Arguments for XAI share concerns with AI's inability to map real causal relationships (London 2019). Humans also lack this ability. We have long used goods from acetaminophen (Ghassemi et al. 2021) and aspirin to lithium (London 2019) without understanding the precise mechanism by which they produce health benefits. Medical practice requires making judgments based on a body of evidence absent full understanding of underlying causes (*id*.). It is unclear why AI should face a distinct explanatory "burden" (Watson et al. 2019) human practitioners need not surmount. Requiring that AI be explainable relies on an undue bias against technology. We would not countenance a world without acetaminophen and its benefits even if its operations could *never* be explained. Losses to human well-being would be too great. Similarly, we should not countenance one where well-validated AI subject to intense scrutiny that could vastly improve human well-being is unavailable because it is unexplainable. This is especially so where strong XAI requirements also offer opportunities for technological error and human misuse of AI data and where strong XAI requirements are unlikely to address underlying problems.[16]

4 The Political Response

Political considerations could still warrant a strong requirement for explainable medical AI. A new argument for XAI in high-stakes settings, like healthcare, focuses on basic liberal-democratic political commitments, not epistemic norms. As Vredenburgh (2022: 210) writes, "[o]paque algorithms threaten to undermine the legitimacy and fairness of the institutions in which they are used." People are owed explanations for decisions that directly impact them. Anyone whose acts impact others' vital interests must justify the acts with reasons those affected cannot reasonably reject as legitimate.[17] Otherwise, they wrong those affected. Governments can only validly constrain subjects' decisions if they offer valid reasons therefor. This plausibly requires understanding for the bases of decisions. Political legitimacy thus requires the use only of AI whose operations and decisions can be explained. Health is a clear vital human interest. So, medical AI must be explainable for the use thereof to maintain basic political legitimacy. The remainder of this work describes and evaluates this 'political response' to XAI-related concerns.

Maclure (2021)'s response to the 'argument from the limitations of human reasoning' is the most complete version of this argument for XAI and responds to some empirical worries in ways that may warrant precisifying relevant epistemic and moral standards. Per Maclure (2021: 427), public reason requires that reasons for decisions that impact individuals be publicly available and acceptable in the sense of being "at least compatible with … a political conception of justice. …

[16] In conversation, Dr. Devin Singh suggests that if medical AI is one of few fields facing a mandatory explainability burden this will drive innovators away from the field. Other sectors will reap the rewards of innovative AI solutions. This posit is outside my scope of inquiry but highlights other possible reasons against strong requirements.

[17] Further to sources above/below, see Kiener (2021)'s similar claims. I use contractualist language amenable to Maclure (2021) and Vredenburgh (2022) to express a general concern with legitimacy. Many liberal-democratic accounts of legitimacy permit similar responses.

[Explainability is then necessary so] automated decisions can be scrutinized and assessed." People, in other words, have a right to know why decisions impacting them were made. Legitimate political decisions must therefore be correct and justifiable. Non-explainable AI fails these conditions. Maclure (422) thus champions a "strong explainability requirement: human organizations … should be legally obliged to demonstrate the capacity to explain and justify the algorithmic decisions … [impacting] the wellbeing, rights, and opportunities of those affected."

Per Maclure (431), human decision-making opacity cannot "vindicate" AI decision-making opacity. Public reason standards are stable across contexts. The argument from the limitations of human reasoning's claim that AI is no worse at meeting them relies on problematic individualist commitments. Reason-giving in politically-salient spheres must instead be examined from an institutional perspective. Large legal and healthcare systems, for example, provide reasons for particular decisions that fill gaps in individual humans' fallible, incomplete, and (sometimes?) opaque reason provision. No analogous systems exist for non-explainable AI; AI researchers cannot "translate a segment of the code into a set of justifications expressible in a natural language" (426). The same may be true of hidden processing nodes in deep learning AI; one cannot translate what one cannot observe. As Maclure (431) writes, "the significance of the social and institutional dimensions of human reasoning is lost from sight" in arguments for non-explainable AI. Institutions provide "more deliberative and transparent" (432) decision procedures (court proceedings, administrative review, etc.) in human cases missing in AI cases.

If Maclure's argument succeeds, it challenges arguments for non-explainable AI beyond the one from human reasoning limitations. If, e.g., using non-explainable AI is illegitimate, remedying technical problems with post-hoc explainability tools is key to securing many benefits of now-opaque AI. While the political response's focus on systems-level production of understanding may not address some XAI advocates' desires, stating that social and institutional processes must provide necessary explanations does not abrogate responsibility to ensure individuals understand them. Per this response, individuals can claim decisions affecting their vital interests are only legitimate where they can be understood. If explanations required at the systems- and individual-levels differ, institutional procedures still produce information that can inform individual decisions. Knowledge gained via rigorous safety review is, e.g., relevant to informed consent.

Maclure's argument is meant to apply in medical settings. Maclure's primary concern is *organizations*' use of non-explainable AI. This covers use by health ministries, healthcare facilities, etc. for rationing or scheduling purposes. It may cover decisions about which AI tools to let on markets or publicly fund. Yet Maclure also discusses the need for explainable treatment or diagnosis decisions and appeals to concepts like informed consent implicating direct provider-patient relationships. This suggests he also desires strong explainability requirements in clinical encounters.[18] Any interpretative issues can, however, be minimized: I will demonstrate that *any* Maclure-style arguments raise issues like those facing other calls for explainable medical AI. Where robust institutional means of ensuring AI safety/efficacy exist, at least some medical AI can be permissibly distributed, used, and even funded absent strong explainability requirements.

[18] Maclure also initially states that public reason does not require understanding of "how the algorithm works" in a case, but only general reasons, and eventually suggests human confirmation of medical AI decisions may meet standards (2021: 426, 435). Yet he faults Geoffrey Hinton for advocating medical AI that does not explain mechanisms by which decisions are made and says reasons for autonomous AI decisions must be transparent for legitimacy and safety reasons, which minimally requires a strong explainability presumption.

5 Issues with the Political Response

The political response has merit but faces many issues with extant arguments for explainable medical AI and raises new problems. Its most plausible form does not respond to relevant challenges. To demonstrate this, I first explain how modified versions of existing institutional administrative procedures can meet public reason standards absent strong explainability requirements. I then clarify why and when clinicians can use some non-explainable AI. I finally detail why strong explainability requirements are not even necessary to fulfill public reason norms for pertinent genuinely political (e.g., rationing) decisions. The arguments jointly show that some non-explainable medical AI can be permissibly distributed, used, and (most likely) funded.

i. Institutional Standards and Public Reason

First grant Maclure's proposed need to examine explainability issues institutionally. Regardless of whether this can apply to particular healthcare provider decisions,[19] public reason standards required for health justice never necessitated full transparency in how medical tools work. They required good reasons for decisions and opportunities to challenge them, which can be and are often provided without tools being explainable. Legal mechanisms for evaluating AI tools present numerous opportunities to assess performance, costs, and reasons for adoption and a framework for assessing accuracy and justifiability, Maclure (2021)'s Rawlsian criteria for assessing political claims. Rather than ban non-explainable health-related AI or even create a defeasible presumption against their availability, I propose strengthening those institutions.[20]

Consider the typical path tools take before implementation in clinical settings in a Western liberal-democracy, like the U.K., U.S.A., or Canada (Minssen et al. 2020; Levine 2020; Homeyer et al. 2021; Flood and Régis 2021). Most medical AI tools begin in clinical trials. Many trials are subject to research requirements from any funding agencies supporting the research and requirements imposed by site (hospital, university, etc.)-specific institutional/research ethics boards. Many are also subject to federal or state regulations, like Health Canada's regulations on research involving humans.[21] Clinical trials consist of several stages. AI tools must meet performance guarantees before going through each step. If performance benchmarks are met, other tests may remain. For instance, a developer whose AI meets the definition of a 'medical device' must pass safety and efficacy review before it will be available on markets. Evidentiary standards at this stage are oft-criticized and some tools escape regulatory scrutiny, but this is also true in non-AI/drugs settings.[22] Such issues apply to all regulated health products, not just AI.

Specific contexts provide further safeguards. In federal countries, state/provincial rules, such as those applying to the regulation of hospitals, determine what goods will be available in a state/province. In publicly-funded healthcare systems, in turn, government

[19] Healthcare providers rarely viewed as political authorities. So, they are not generally subjects of constitutional claims in countries where the constitution only applies to government actors. E.g., *Eldridge v British Columbia (Attorney General)*, [1997] 3 SCR 624. Claimed authority is epistemic or practical. However, the response could, again, have individual-level implications.

[20] If Maclure agrees, consider this a friendly amendment. It would not maintain the political response as a critique of non-explainable medical AI. And Maclure is not the sole strong explainability requirement proponent.

[21] Many institutions are also establishing local AI-specific research standards (Flood and Régis 2021).

[22] E.g., Price (2017) (on the U.S.A.); Flood and Régis (2021) (on Canada).

decision-makers have mechanisms for deciding which goods are available. For instance, Canadian provinces' decisions about what to add to the public medical formulary (viz., goods available in a public system) are often informed by additional studies (e.g., health technology assessments) using quality, efficiency, cost, and other relevant metrics (Flood and Régis 2021). One must follow every step in such an institutional process before providers get a good. Then, individual providers must make decisions consistent with professional practice guidelines and private law care standards. Each step in the process offers information that can help 'justify' using a good and engages with values relevant to whether the decision/action comports with public reason. If each step has issues in particular states, they still jointly provide a framework for assessing if using a good is justifiable. They suffice for assessing otherwise-opaque human decisions and should for AI.

'Institutional' considerations, then, demonstrate how extant bodies that assess medical tool development and use can assess the justifiability of non-explainable AI distribution and use. This is true for model- and clinician-centered evaluative frameworks. Some existing mechanisms for assessing medical tools admittedly present difficulties. Not every AI tool faces each stage of scrutiny above. But regulatory imperfections cannot vindicate strong explainability requirements. For instance, safety and efficacy reviews of drugs and medical devices are often criticized for their low evidentiary standards (Da Silva et al. 2022). Yet standards are not low enough to render use of all relevant goods unjustified. Other stages of review provide safeguards against digital snake-oil. If one does not trust humans to serve as proper safeguards against misuse, it is hard to see why this should favour continuing to rely on human providers' decisions over those of non-explainable AI. If the problem at hand is poor human performance, more human action is an unusual solution. Moreover, evidentiary problems here likely apply to all health products; they accordingly do not support a strong explainability requirement for medical AI alone. Those interested in evidence should look to AI to fill gaps in human knowledge. Those interested in 'correct' decisions should not accept explainability/accuracy trade-offs too quickly.

Extant institutions require modifications to account for AI-related issues but many issues can plausibly be addressed without strong explainability requirements. For example, Ratti and Graves (2022) suggest regulators should require explanations of how tools were built, including documentation of and justification for "technical choices … made in designing" ML tools. They note that current standards propose innovators should "provide a long list of technical requirements and specifications, spelled out as neutral, step-by-step recipes … [and] reasons why the technical choices made are best and result in the overall effect." While they admit this problem would address the opacity of procedures to train algorithms, not algorithms themselves, this could address many underlying concerns. For another, as I discuss in some detail in prior work with Colleen M. Flood and Mathew Herder (Da Silva et al. 2022), regulators in Western liberal-democracies are alive to unique problems by truly adaptive ML tools, like difficulties identifying problems ex-ante, and offer principles to guide ongoing regulatory reforms for AI-enabled medical devices. My collaborators and I promote strong mechanisms for post-market scrutiny, including regular third-party audits, and argue they must not come at the expense of pre-market scrutiny. Indeed, we contend, evidentiary standards for pre-market review of adaptive ML-learning enabled devices should be higher short-term; as noted above, ML 'effectiveness' standards are in flux and regulators should be cautious while they develop. We further note that regulators expanded the understanding of safety in the past and should do so for AI to address bias-related safety concerns. Such proposals and Tschandl et al. (2020:1232)'s call for real-world testing of AI could also address concerns about link uncertainty if they require evidence sufficient to establish causal relations. The key is that plausible reforms do not require XAI.

Direct-to-consumer AI tools may not face all stages of review above but face some stages deemed acceptable for non-AI consumer goods and are subject to products liability norms that constrain use. Institutional parallels thus remain. Problems with extant institutional safeguards are, again, not clearly issues with non-explainable AI. They point to general needs to improve mechanisms for ensuring that healthcare decisions are justifiable across the spectrum of care. Where any problems also apply to drugs and other devices, one must improve the system. Strong explainability requirements leave extant worries in place and bar use of beneficial tools.

Whether administrative decision-making itself is legitimate is largely beyond my scope of inquiry.[23] However, administrative decisions are subject to review and otherwise meet basic public reason standards. Modifications above would improve matters. If they do not vindicate all administrative decision-making, that would not undermine my proposal. Arguments for administrative bodies' legitimate regulation of other healthcare goods plausibly apply to AI, regulation of AI can meet public reason standards viewed as legitimate elsewhere, and regulators can help ensure decisions about AI use meet them too. Non-explainable AI tools meet justifiability standards and limitations of explainability mechanisms suggest many *better* meet correctness standards. Barring them is thus unmotivated. If some AI opacity makes checks more difficult, mechanisms above/below validates ex-ante and -post safety and efficacy. There may be cases where one AI error has severe consequences. But permitting lower-risk non-explainable AI is advisable in more common cases where humans fare no better and safeguards on AI exist.

Explainability tools may be useful elements of an auditing process, but a tool's ease of submission to XAI processes should not determine of whether it can be licenced and used if it passes robust alternative auditing processes. SHAP and LIME, for example, may be useful for model auditing even if they do not fully illuminate AI operations.[24] Knowing which inputs are most likely to contribute to outcomes indicates plausible bases for AI decision-making and can help evaluate the (probable) underlying mechanism. This can help address concerns a given tool has only proved safe and effective in existing studies due to overfitting by providing an additional means of auditing decisions beyond brute outcome results. Yet SHAP values may not always be determinative of whether we should permit an AI tool's use. Consider a tool that consistently meets high standards for safety and efficacy when tested on diverse populations in diverse settings. Now imagine that it continues to meet and even exceed those standards when provided with distinct datasets that differ in important ways from initial training data. Further imagine that its developers agreed to additional auditing whenever the AI tool fell outside a performance window and that the tool did not do so for a lengthy period of time. Following the tool's recommendations would lead to better long-term health outcomes than following those of the average healthcare provider prior to its development. If submitting that tool to a SHAP value analysis produced odd stories about which inputs contributed to its recommendations, I see little reason to think one should take it off the market. One should, following Sullivan (2022), likely use that outcome as a starting point for further scientific research. The facially-bizarre result may warrant closer scrutiny of underlying empirical data and further auditing using other tools. But the continued sale/use of the tool under these circumstances strikes me as desirable, particularly given the fact that an XAI tool could have the same or worse performance issues as any AI.

[23] I thank Kate Vredenbergh for raising this apt issue during discussion of a much earlier draft at the Frankfurt School of Finance and Management's The Philosophy of Data Science: Data Science Governance conference.

[24] Compare note 9.

ii. Clinical Encounters and Public Reason

One may then seek to focus on the importance of explainability in clinical settings. Doing so does not clearly aid Maclure et al. Clinicians must justify their actions, which provides some support for explainability requirements (London 2019). Public reason may be a useful framework for judging clinical decisions. However, the institutional dimensions of existing clinical decision-making above also support a parallel 'political' argument from the limitations of human reasoning. Justifiable human clinical decision-making need not be transparent nor informed by perfect evidence. We expect it to conform to our best safety and efficacy standards and establish institutional safeguards to limit risks of unsafe or ineffective care and inform our standards for what decisions are 'justified.' It is unclear why we should expect more of AI.

Clinical decisions are, in short, already informed by a series of institutional processes that provide information clinicians can and do use to justify decisions, from safety and efficacy review to professional guidelines. The clinical standard of care is informed by those proceedings and sets evidentiary standards justified decisions must meet. These mechanisms are widely viewed as providing an acceptable, if imperfect, framework for making justified decisions about drugs or non-software-based medical devices absent full understanding of the mechanisms under which goods work or perfect knowledge of their accuracy in a case. This too is a basic outline of how medical practice operates in modern administrative states. It is accordingly unclear how an institutional perspective undermines the argument from the limitations of human reasoning.

Clinicians' legal and moral duties of explanation in informed consent contexts (e.g., Froomkin et al. 2019; Kiener 2021) also do not establish strong explainability requirements. Extant norms do not require explanations of the mechanisms by which options will work that would bar lithium prescriptions (*id.*). Many states require disclosing potential risks and benefits of recommendations and alternatives but not the fact that clinicians do not know how a healthcare good works (Cohen 2020; Froomkin et al. 2019). The fact of AI opacity should plausibly be disclosed, as when dealing with opaque novel drugs. So too should any odd results identified in auditing processes, including any XAI-based ones. But that need not bar non-explainable AI, particularly where many patients want to benefit from the most accurate tools, AI or otherwise.

Informed consent processes, then, are yet another institutional process for minimizing the chances of unsafe AI use but do not support strong explainability requirements. Any patient-centered explanatory requirement should not require knowledge of, e.g., how AI operates.

A recent concern about AI paternalism (Luxton 2022; Kühler 2022; Diaz Milian and Bhattacharyya 2023) raises regulatory challenges but also does not support strong explainability requirements. Concerns about AI paternalism identify three related but severable issues (i) AI tools can unduly and covertly influence persons for their own sake in objectionable ways that cannot be attributed to a human (Kühler 2022), (ii) AI recommendations can be given undue priority in decision-making (Luxton 2022; Diaz Milian and Bhattacharyya 2023), and (iii) AI decisions can fail to account for patient preferences (*id.*). These concerns require responses that are alive to how AI is regulated and integrated into clinical encounters but do not favour strong XAI requirements.

The health applications undergirding (i) (Kühler 2022) often avoid regulatory scrutiny (Da Silva et al. 2022). The risk underlines the need for regulation. But XAI is not clearly necessary or sufficient to address the basic problem. AI-provider-patient

interactions undergirding (ii) and (iii), in turn, point to the need for a regulatory process that ensures AI and providers provide enough information to patients so they can make decisions and that both respect patient choices. AI introduces additional loci for recommendations that complicates the informed consent process. However, the need to respect patient preferences is orthogonal to explainability questions. And the level of information patients require to make decisions once again does not support strong explainability requirements. Paternalism charges highlight the need to look at the entire decision-making ecosystem when deciding whether to use AI or follow its decisions. Anti-paternalist calls not to categorically prioritize AI recommendations (Luxton 2022; Diaz Milian and Bhattacharyya 2023) and consider building patient preferences into the original design (Luxton 2022) have merit. But paternalism charges only support strong explainability requirements if persons cannot make free and informed choices absent knowledge of AI operations. Such knowledge is not required for other informed choices and should not be here.

AI thus complicates healthcare provision, including informed consent processes, and underlines the importance of attending to the contexts for AI use. Safety and efficacy standards and clinical standards of care should attend to these complications. Scrutiny of how AI will play a role in real-world shared decision-making is necessary. But resolving these general issues does not require that regulators, clinicians, or patients understand how all safe and effective AI operates.

iii. Genuinely Political Decisions and Public Reason

One may then be tempted to focus on the public dimension of the political response. Public reason norms more directly apply to governments. The response could plausibly only apply to public decisions about whether to permit or fund medical AI use or AI-based public rationing.[25] This understandable move nonetheless falters. Existing mechanisms for deciding what can be permitted or funded again do not require explainability. Nearby claims that public decisions must be transparent about their underlying reasons then simply accept that we need mechanisms like those already in place (with modifications). This argument also raises questions about whether appeals to public reason even 'respond' to purported issues with non-explainable medical AI. Explainability skeptics (e.g., London 2019; Ghassemi et al. 2021; Babic et al. 2021) claim that permitting the development and use of non-explainable AI medical can be justified. Stating that only non-explainable that meets justifiability thresholds is acceptable grants their point.

Limiting the political response to rationing decisions is likewise problematic. Non-explainable AI appears to meet the most commonly used standard for legitimate medical rationing decisions, Daniels and Sabin (2002)'s accountability for reasonableness framework. That framework requires (a) public decisions that are (b) relevant to healthcare decision-making and relevant stakeholders can be expected to accept and (c) opportunities to review and appeal decisions. It also requires (d) regulations that guarantee (a)-(c). (a)-(d) are consistent non-explainable AI use. If frameworks for rationing healthcare largely fulfill them and permit and even publicly fund using other opaque medical products with proven efficacy, few find this problematic. And if existing

[25] This would explain Maclure (2021: 426)'s focus on 'organizations' and Ontario (2022)'s focus on public AI uses. Vredenburgh (2022) likewise frames her arguments around 'organizations' (but includes private ones).

regulatory reviews fail to provide adequate reasons for decisions or appellate proceedings, this is a problem. But it is not a problem *with non-explainable AI*.

If the political response only highlights the need for justified reasons for using medical AI, it does not undermine the case for public decisions to permit or fund non-explainable medical AI use (or clinical decisions to use them). The response thus does not vindicate a strong requirement for XAI. It supports a requirement to explain when and why non-explainable medical AI use is justified. Existing frameworks provide methods for analyzing the case for medical AI that are widely accepted for non-explainable drug use, for example. Problems with extant processes justify reforming institutions, not prohibiting non-explainable medical AI. Barring our most accurate AI in the name of problems that do uniquely apply to AI is unwise, especially given possible unintended consequences. This not an assertion that "whatever heals is right" (Herzog 2022). It is an *argument* for why public reason does not require barring many healing goods.

Claims that a human-in-the-loop must confirm AI decisions (Maclure 2021: 435) are, finally, largely orthogonal to present debates. Studies above provide mixed evidence on whether human-XAI interactions improve performance. At best, whether human-AI interactions are preferable to alternatives depend on other institutional design questions (Parasuraman and Wickens 2008). I above granted the need to attend to broader context of use when evaluating the permissibility of any health product. But one need not settle questions about the value of a human-in-the-loop to address the political response or alternatives. Debates about AI explainability and the need for a human-in-the-loop are analytically distinct. Technical and philosophical issues with explainability requirements appear where humans are present and humans can 'check' AI decision-making absent AI explainability requirements, as in drug and other device settings.

6 Conclusion

Requiring explainable medical AI has significant costs that are not worth it where they are unlikely to bring about significant corresponding benefits. Prohibiting well-vetted tools with a track record of effective treatment recommendations for what has been treatment-resistant depression or identifying cancers requiring early interventions earlier, better, and cheaper due to desires to understand how they work is untenable. Requiring explanations of how other tools work or how clinicians reach decisions is unwarranted. Established institutional frameworks clarify which decisions are warranted when and make it far less likely that clinicians will use unsafe tools or use tools in unsafe ways. Largely-acceptable systems for assessing the use of non-explainable medical decisions that can apply equally to non-explainable AI. They may need modifications to address AI but that does not alter the balance of reasons supporting use of non-explainable medical AI. Most problems support law and policy reform regardless of whether one permits non-explainable medical AI use, rather than providing distinct reasons against use.

Appeals to public reason to assess whether licencing, use, or funding of medical AI is justified highlight the need to justify all medical decisions that impact persons' vital interests but do not address philosophical and technical problems with arguments for strong XAI requirements. Claims that public reason standards require

the exclusive use of explainable medical AI face similar issues as other arguments therefor. Institutional considerations motivating Maclure's most forceful arguments instead demonstrate that political arguments for explainable medical AI face a parallel argument from the limitations of the human mind and likely raise arguments from accuracy and unintended consequences. One may justifiably seek to improve AI explainability for given ends, like fostering patient trust. But strong XAI requirements raise issues and public reason does not require them where strong institutional review exists.

Other claimed reasons for strong XAI requirements do not survive scrutiny. Requirements could, e.g., set epistemic standards individuals must meet when deciding what to do, which may map responsibility categories; this could help us regulate non-explainable AI (Yoon et al. 2022). Yet responsibility for opaque decisions could, in principle, be justifiably apportioned. Doing so will be difficult and requires scrutinizing whether relevant harms are best analyzed through product liability law or medical malpractice law and the relationship(s) between those domains. But knowledge of outcome measures for non-explainable AI can do some work. Provider should, e.g., be liable for using a tool on populations for whom it has not been validated, regardless of whether anyone understands the mechanism by which it works for some populations and not others. This is how we assess liability for similar harms. Private law specialists are, moreover, tailoring standards for medical AI. For instance, Price II et al. (2022) posit provider duties to assess the quality of the non-explainable AI developers and ensure external validation of a tool occurred to avoid liability. Warning labels can also play a role in setting proper standards, as they do elsewhere (Da Silva et al. 2022).[26] Contra Yoon et al. (2022), existing legal standards can and do attend to cases where full explanations are unavailable. Requiring XAI would, moreover, be problematic for reasons above even if it were always easier to apportion liability for XAI use.

Valid concerns about how to translate reasons into code and back again or identify opaque AI-related problems ex-ante or solutions ex-post do not fully undermine arguments for using at least low-risk non-explainable AI. Barring valuable non-explainable AI when risks are low, explainability tools compound risks of error, and we accept other forms of opaque medical decision-making is undermotivated. One should not give up on non-explainable AI's enormous transformative potential due to bare desires for human care. When tempted to do so for political justice reasons, one should recall humans' many opportunities to weigh reasons for and against using a tool in existing administrative frameworks.[27]

[26] For a fascinating discussion of this possibility as well as some of its challenges (including some related to black-box issues), which was released after this text was complete, see Gerke (2023). It does not change the present point.

[27] I thank anonymous reviewers, Hannah Da Silva, Jocelyn Maclure, Dr. Devin Singh, and Daniel Weinstock, and audiences at the Southampton Ethics Centre and the Frankfurt School of Finance and Management's The Philosophy of Data Science: Data Science Governance conference for feedback on prior drafts. This piece originated as lecture notes for an intensive on the law and ethics of health-related AI at the University of Ottawa. I also thank my students and guests (Dr. Singh, Melissa McCradden, and Marc Lamoureux) for related discussions. During the course, I was the Alex Trebek Post-Doctoral Fellow in AI and Health Care and part of the Machine M.D. project at Ottawa. I thank the Alex Trebek Forum for Dialogue and Canadian Institutes of Health Research for funding my job as well as the Machine M.D. team for their support.

Declarations

Competing Interests The author declares no competing interests.

References

Amann J et al (2020) Explainability for artificial intelligence in healthcare. BMC Med Inform Decis Mak 20:210
Arrieta AB et al (2020) Explainable Artificial Intelligence (XAI). Inf Fusion 58:82–115
Babic B et al (2021) Beware explanations from AI in health care. Science 373(6552):284–286
Balagopalan A et al (2022) The road to explainability is paved with bias. 2022 ACM Conference on Fairness, Accountability, and Transparency: 1194–1206
Benjamin R (2019) Race after technology. Polity, Cambridge
Bringsjord S, Govindarajulu NS (2018) Artificial intelligence. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/artificial-intelligence/
Cohen IG (2020) Informed consent and medical artificial intelligence. Georgetown LJ 108:1425–1469
Da Silva M et al (2022) Regulation of health-related artificial intelligence in medical devices: the Canadian story. UBCLR 55(3):635–682
Daniels N, Sabin JE (2002) Setting limits fairly. Oxford UP, Oxford
Diaz Milian R, Bhattacharyya A (2023) Artificial intelligence paternalism. J Med Ethics 49:183–184
Flood CM, Régis C (2021) AI & Health Law in Canada. In: Bariteau-Martin F, Scassa T (eds) Artificial intelligence and the law in Canada. LexisNexis
Froomkin AM et al (2019) When AIs outperform doctors. Ariz LR 61:33–99
Gerke S (2023) Nutrition facts labels' for artificial intelligence/machine learning-based medical devices: the urgent need for labeling standards. George Washington LR 79:91–163
Ghassemi M et al (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2:e745-750
Herzog C (2022) On the ethical and epistemological utility of explicable AI in medicine. Philos Technol 35(2):50
Homeyer A et al (2021) Artificial intelligence in pathology. J Pathol Inform 12:1–13
Jacobs M et al (2021) How machine-learning recommendations influence clinician treatment selections. Transl Psychiatry 11:108
Johnson DG (2021) Algorithmic accountability in the making. Soc Philos Policy 28(2):111–127
Kiener M (2021) Artificial intelligence in medicine and the disclosure of risks. AI Soc 36:705–713
Kühler M (2022) Exploring the phenomenon and ethical issues of AI paternalism in health apps. Bioethics 36(1):194–200
Levine HR (2020) Anticipating regulatory reform. Seton Hall LR 50:805–826
Lindsell CJ et al (2020) Action-informed artificial intelligence. JAMA 323(21):2141–2142
London AJ (2019) Artificial intelligence and black-box medical decisions. Hastings Cent Rep 49(1):15–20
Lundberg SM et al (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67
Luxton DD (2022) AI decision-support: a dystopian future of machine paternalism? J Med Ethics 48:232–233
Maclure J (2021) AI, explainability and public reason. Mind Mach 31(3):421–438
Minssen T et al (2020) Regulatory response to medical machine learning. J Law Biosci 7(1):1–18
Obermeyer Z et al (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453
OECD (2019) Recommendation of the Council on Artificial Intelligence (OECD Legal Instruments). OECD/LEGAL/O449
Ontario (2022) Beta principles for the ethical use of AI and data enhanced technologies in Ontario. https://www.ontario.ca/page/beta-principles-ethical-use-ai-and-data-enhanced-technologies-ontario

Panch T et al (2019) Artificial intelligence and algorithmic bias. J Glob Health 9(2):020318
Parasuraman R, Wickens CD (2008) Humans: still vital after all these years of automation. Hum Factors 50(3):511–520
Pierce R et al (2022) A riddle, wrapped in a mystery, inside an enigma. Bioethics 36(2):113–120
Pigoni A et al (2019) Can machine learning help us in dealing with treatment resistant depression? J Affect Disord 259:21–26
Poursabzi-Sangdeh F et al (2021) Manipulating and measuring model interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 237:1–52
Price WN II (2015) Black-box medicine. Harv JL Tech 28:419–467
Price WN II (2017) Regulating black-box medicine. Mich LR 116(3):421–474
Price WN II (2018) Medical malpractice and black-box medicine. In: Cohen IG et al (eds) Big data, health law and bioethics. Cambridge UP
Price WN II et al (2022) New innovation models in medical AI. Wash ULR 99:1121
Ratti E (2022) Integrating artificial intelligence in scientific practice. Philos Technol 35:58
Ratti E, Graves M (2022) Explainable machine learning practices. AI Ethics 2:801–814
Rawls J (1993) Political liberalism. Columbia UP
Roy A et al (2020) A machine learning approach predicts future risk to suicidal ideation from social media data. npj Digit Med 3:78
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Russell SJ, Norvig P (2021) Artificial intelligence, 4th edn. Pearson, Harlow
Simkute A et al (2021) Explainability for experts. J Responsible Technol 7–8:100017
Sullivan E (2022) Understanding from machine learning models. Br J Philos Sci 73(1):109–133
Tomsett R et al (2018) Interpretable to Whom? ArXiv 1806:07552
Topol E (2019a) Deep medicine. Basic Books, New York
Topol E (2019b) High-performance medicine. Nat Med 25:44–56
Tschandl P et al (2020) Human-computer collaboration for skin cancer recognition. Nat Med 26:1229–1234
Ursin F et al (2022) Explicability of artificial intelligence in radiology. Bioethics 36(2):143–153
Vredenburgh K (2022) The right to explanation. J Polit Philos 30(2):209–229
Watson D (2021) Explaining black box algorithms. DPhil Thesis, Oxford University, Oxford
Watson DS (2022a) Conceptual challenges for interpretable machine learning. Synthese 200:65
Watson DS (2022b) Interpretable machine learning for genomics. Hum Genet 141:1499–1513
Watson DS, Floridi L (2021) The explanation game. Synthese 198:9211–9242
Watson DS et al (2019) Clinical applications of machine learning algorithms. BMJ 364:l886
Watson DS et al (2022) Local explanations via necessity and sufficiency. Mind Mach 32:185–218
Yap M et al (2021) Verifying explainability of a deep learning issue classifier trained on RNA-seq data. Sci Rep 11:2641
Yoon CH et al (2022) Machine learning in medicine. J Med Ethics 48:581–585
Zednik C (2021) Solving the black box problem. Philos Technol 34:265–288
Zednik C, Boelsen H (2022) Scientific exploration and explainable artificial intelligence. Mind Mach 32:219–239
Zerilli J (2022) Explaining machine learning decisions. Philos Sci 89:1–19
Zimmermann A et al (2022) The political philosophy of data and AI. Can J Philos 52:1–5