Lessons from Clinical Communications for Explainable AI

Alka V. Menon[1], Zahra Abba Omar[1], Nadia Nahar[2], Xenophon Papademetris[1], Lynn E. Fiellin[3], Christian Kästner[2]

[1]Yale University, [2]Carnegie Mellon University, [3]Dartmouth University

Abstract

One of the major challenges in the use of opaque, complex AI models is the need or desire to provide an explanation to the end-user (and other stakeholders) as to how the system arrived at the answer it did. While there is significant research in the development of explainability techniques for AI, the question remains as to who needs an explanation, what an explanation consists of, and how to communicate this to a lay user who lacks direct expertise in the area. In this position paper, an interdisciplinary team of researchers argue that the example of clinical communications offers lessons to those interested in improving the transparency and interpretability of AI systems. We identify five lessons from clinical communications: (1) offering explanations for AI systems and disclosure of their use recognizes the dignity of those using and impacted by it; (2) AI explanations can be productively targeted rather than totally comprehensive; (3) AI explanations can be enforced through codified rules but also norms, guided by core values; (4) what constitutes a "good" AI explanation will require repeated updating due to changes in technology and social expectations; (5) AI explanations will have impacts beyond defining any one AI system, shaping and being shaped by broader perceptions of AI. We review the history, debates and consequences surrounding the institutionalization of one type of clinical communication, informed consent, in order to illustrate the challenges and opportunities that may await attempts to offer explanations of opaque AI models. We highlight takeaways and implications for computer scientists and policymakers in the context of growing concerns and moves toward AI governance.

Introduction

As AI models become more commonly used, and most of them are inscrutable even to their developers, explainability has become an increasingly important consideration. Recent policy proposals for regulating AI, including the EU Artificial Intelligence (EU AI) Act (European Parliament 2024), the White House Blueprint for an AI Bill of Rights (U. S. White House 2022), the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order 14110 2023), and the NIST Risk Management Framework for Generative AI (National Institute of Standards and Technology 2024) lay out provisions for notice and explanation for automated systems. But it remains unclear and contested what constitutes an adequate explanation for an AI system, particularly across different use cases.

There is disagreement on what explanations for AI systems should look like, what they are for, how they work, and what effect, positive or negative, they have on how humans make decisions (Ribera and Lapedriza García 2019; Gilpin et al. 2018) using such systems. In practice, there are many papers and tools focused on explainability (and the wider notion of transparency), but they are used primarily by developers for debugging their models (Bhatt et al. 2020; Kaur et al. 2020). Under the label *human-centered explainable AI*, there is also a substantial body of research on which kind of explanations might be suitable for specific *end-user* tasks and human-in-the-loop oversight designs, e.g., (Ehsan and Riedl 2020; Vera Liao and Varshney 2021; Dedikov 2023; Rong et al. 2022; Ehsan et al. 2021; Luria 2023; Panigutti et al. 2023; Jia et al. 2023; Stumpf, Bussone, and O'sullivan 2016), though empirical evidence of effectiveness is mixed (Rong et al. 2022) and deployments of end-user explanations are still relatively rare. Guidance for how to design effective explanations is limited – developers have a large tool box (Molnar 2022) and high-level suggestions (Yildirim et al. 2023), but it is often unclear how to approach explanations.

Stepping back from the concrete, practical challenges of setting guidelines and standards for explainability, we find it helpful to find an analogue for the task of communicating complex, technical information to non-experts. Our interdisciplinary research team found the norms associated with *clinical communication*s to be especially illuminating of the benefits and challenges of this task. This position paper comes from a multi-year research collaboration between experts in the fields of computer science, sociology, engineering, and medicine. Though we did not initially share a common vocabulary, we found certain examples to be critical for establishing a shared understanding of concrete policy solutions and insights, which then paved the way for empirical research on explainability and policy. These examples served as boundary objects for our research team, crossing the boundaries of our respective fields (Star and Griesemer 1989; Carlile 2002, 2004). Clinical communications generally, and informed consent specifically, helped us unlock how we could operationalize transparency and explainabil-

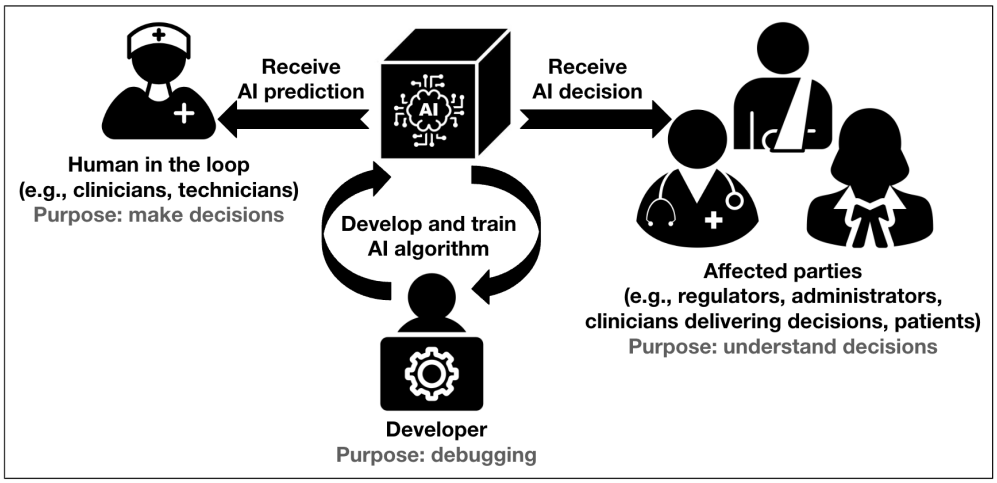

Figure 1: Explanations could be offered at different points in the process and to fulfill different needs

ity. Conventions for clinical communications are well established but have been much debated over time, making them a rich source for previewing debates to come over explainability for AI systems.

In clinical settings, there is an *asymmetry of knowledge* between physicians and patients, just as there is between the builders of an AI system and users. Physicians can offer some explanation of how they came to a diagnosis or treatment selection, but a detailed explanation may be *too technical or confusing* for patients. While physicians may consciously know the components leading to a diagnosis, they may sometimes be faced with gaps and need to rely on their intuition. Physicians' explanations of how they arrived at a treatment plan or course of action thus may have an element of *justification*, that is, a post-hoc narrative that aligns the decision with a valid reasoning sequence. In practice, physicians may need to give *different degrees of detail* when explaining a diagnosis to colleagues in healthcare compared with patients. And different explanations might be required for patients depending on their intended aim, with physicians demonstrating concern about overwhelming patients (and their families) with too much information but wanting to provide enough information for patients to ask questions and engage in their care. Physicians are often trying to persuade a patient of a recommended treatment, based on medical knowledge and best practices. They present the patient with the potential benefits but also the risks of a given course of action.

In this position paper, we describe how the informed consent process from clinical communications can be used to anticipate the challenges and benefits of providing explanations for opaque AI systems. We trace parallels not to suggest that the informed consent process is an exemplar to be exactly replicated, but to offer recommendations for computer scientists and policymakers as to how these may influence their work going forward. Our goal in this position paper is to inform current debates in policy and computer science about whether and how to require explanations for AI systems.

In what follows, we outline a brief history of informed consent and how it works in clinical settings, continuing on into lessons that we can learn from them. We argue that an exploration of the history of informed consent and the debates it has engendered can help surface concerns, and allow us to make recommendations and predictions about regulatory and governance strategies for AI.

Explainability in AI

Machine learning (ML) is a field of artificial intelligence (AI) that inductively learns functions (called *models*, and in policy discussions often just *algorithms*) from data to make predictions.[1] "Machine learned" models can have a simple human-interpretable structure such as if-then-else decision trees and linear models. However, in practice, models with complex inner structures (e.g., trillions of parameters in a neural network, such as in the case of ChatGPT v4) are often used that are inscrutable (impossible to fully understand) even to their developers. Machine-learned models can be effective if they provide mostly useful predictions, even if we do not know how they work or what they have learned. But opacity can also lead to problems if the model has learned irrelevant and misleading rules and shortcuts (e.g., distinguishing between a wolf and a husky based on an image's background (Joshi, Agarwal, and Lakkaraju 2021) or learned unjust discriminatory rules (e.g., reproducing gender bias in the training data (Latif, Zhai, and Liu 2023; Larrazabal et al. 2020).

There are many kinds of information we might want to know about a model, such as: How does the model work generally (e.g., what features is it sensitive to)? How did the model arrive at a specific prediction (e.g., diagnosing cancer from a medical image)? What factors were most important

[1] In this paper, we use the more general term "AI," following policy discourse. Our discussions of explainability primarily focus on machine learning enabled software systems.

Table 1: Summary of key lessons from clinical communications for computer scientists and policymakers

Lessons from informed consent	Advice to computer scientists	Advice to policymakers
Lesson 1. Clinical communications are provided to advance patient autonomy.	AI explanations to end users may have intrinsic value.	Specify a purpose for AI explanations.
Lesson 2. Information provided to patients is purposefully targeted in a way that ultimately benefits patients.	Instead of one comprehensive explanation, consider offering multiple targeted explanations for different users.	The degree of detail required in AI explanation may vary based on risk.
Lesson 3. Clinical communications are governed by both rules and norms.	Discuss and develop professional norms around AI explainability.	Any rules raise the possibility of bare-minimum compliance. Ask for evidence.
Lesson 4. "Good" clinical communications change over time.	Consider values and principles first in constructing AI explanations, and quickly changing explainability techniques second.	Solicit opportunities for public feedback. Plan for regular updates of rules for AI explanations.
Lesson 5. Clinical communications can have broader effects on the public's trust in doctors and medicine.	Offer explanations for AI routinely, not just when something goes wrong.	Be open and transparent with the public about how policymakers evaluate AI, including AI explanations.

in arriving at the prediction? What data was used for that prediction? How was the model trained and what data was used to train the model? These questions are usually easy to answer for human-interpretable models like shallow decision trees and sparse linear models. For complex, opaque models, however, answers can only be approximated. Many model explainability techniques and tools are available for such approximations (Molnar 2022), explaining the model as a whole, individual predictions, or the data involved.

There are debates about when and whether opaque AI models can or should be used. In computer science, the dominant view is that opaque models are more expressive and generally better (Bell et al. 2022) though see the work of Cynthia Rudin (Rudin 2019) for a critique of this view. Social science and legal scholars have tended to raise concerns about fairness and trustworthiness of opaque models (Broussard 2023; Benjamin 2019; Selbst and Barocas 2018; O'Neil 2016). Both sides recognize, however, that explainability requires making trade offs between qualities such as completeness, accuracy, and transparency. It is worth noting, however, that the most attention-grabbing recent developments in AI, including generative AI and large language models (e.g., ChatGPT, LLama), lie at the most opaque end of the spectrum.

Overall, explanations for *AI systems* – that is, software systems that use machine-learned models for predictions and decisions – can have different purposes, as depicted in Figure 1. Explanations can be used for general understanding and debugging of models, can be used to foster joint human-AI decision making, or can be used to justify decisions to affected parties and to provide information for possibly appealing the decision (Bhatt et al. 2020; Jacovi et al. 2021; Selbst and Barocas 2018). The purpose of an explanation is not always clearly articulated, and one single explanation may not adequately suffice for all of the above purposes. Our goal here is not to point to one perspective or another, but to highlight the strengths, weaknesses, and consequences of providing different kinds of explanations for AI systems by drawing on the well-established example of informed consent in the field of medicine.

Informed Consent in Clinical Communications

In the clinic, as opposed to in research settings, informed consent is a framework for healthcare providers to convey information about a procedure or treatment to patients.[2] As outlined by the American Medical Association (AMA), to obtain informed consent for care from patients, physicians should convey the risks and benefits of treatment to a patient, who has been determined to be a competent decision-maker, in middle-school-level language with minimal jargon. As part of the conversation, they should give alternative treatment options and invite questions from patients. Patients ultimately have the opportunity to refuse treatment (American Medical Association Opinion 2023b). In clinical care, as distinct from its use in research, informed consent is typically a conversation that a physician documents in the patient's electronic health record. For procedures, however, such as radiologic or surgical procedures, the informed consent process is often documented by a multi-page document signed by both parties. For patients who are unable to legally provide consent (due to young age or incapacity), U.S. guidelines require physicians to obtain the patient's assent, alongside the informed consent for their care from a le-

[2]We focus on the use of informed consent for clinical care rather than in human subjects research, as people cannot obtain clinical care without engaging in informed consent practices (though they can generally obtain care while refusing to participate in research). However, we draw on literature in bioethics and informed consent across these cases for insights as relevant. In this paper, we use the term "physician" instead of healthcare providers because the specific guidelines we discuss apply to medical doctors, though other healthcare providers (e.g., dentists, nurses) adopt similar practices.

gal guardian. That is, even when patients are not considered legal agents, they must be consulted and involved in their own care through a targeted conversation about their treatment. Informed consent opens a channel between physicians and patients and sets a baseline level of information about treatment options. The goal of informed consent is to provide individual patients with information about a care plan, directly involving them in decision-making about their treatment. It is a process that has become ubiquitous, aiming to ensure patient understanding as well as acceptance.

The idea that physicians have obligations around disclosure to patients has been a longstanding ethical ideal in medicine. Clinical and medical ethics dating to Hippocrates have included the mandate for physician-caregivers to be truthful to patients (Beauchamp 2011). Historically, physicians were expected to pursue the best interests of a patient, with norms varying as to how much information physicians disclosed to patients about their treatment. The requirement for physicians to tell patients the truth *and* obtain their informed consent for treatment is now traced to the principle of patient autonomy in clinical ethics (Varkey 2021). U.S. court cases established the primacy of patient autonomy in the early 20th century, beginning with *Mohr v. Williams* in 1905 (Bazzano, Durant, and Brantley 2021). A 1957 U.S. Supreme Court case established "informed consent" as we know it today, adding the requirement that patients understand the information conveyed to them. Further legal decisions established (1) what was necessary to demonstrate patient understanding and (2) compliance with institutional and medical norms (Beauchamp 2011). The 1979 Belmont Report, commissioned by the U.S. National Research Act to establish ethical principles around the intersection of clinical care and medical research, has also been influential in shaping informed consent procedures. The Belmont Report linked three ethical principles, respect for persons, beneficence, and justice, to regulatory requirements such as documentation of informed consent, risk-benefit calculations, and recruiting research subjects without discriminating (Belmont Report 1979). Today, informed consent is a legal requirement as well as a norm.

In contrast to informed consent for human subject research participation, informed consent in clinical settings is not governed by federal oversight, but at the organizational and professional levels. Hospitals and clinics have generated standardized language about interventions and procedures, offering templates for communicating the risks of a treatment, especially for experimental drugs or procedures. This means that documentation for informed consent is set by institutional committees and can vary across healthcare settings for the same procedures.

Thus, informed consent provides a culturally negotiated script for navigating the imbalance of information and power between physicians and patients. Informed consent forms are standardized, but the process (and art) of obtaining informed consent is taught by physicians to their medical trainees through an informal process of apprenticeship. These norms—of how to build rapport with patients, of how to go beyond what's on the page of an informed consent document—allow some flexibility to adapt to different patient-provider relationships and diagnosis and treatment conditions. As medicine has aspired to "shared decision-making" between patients and physicians, it has become more important to ensure that clinical communication is effective.[3]

Clinical communications like informed consent are not a perfect match for explanations for AI systems (hereafter referred to as "AI explanations"). Importantly, informed consent is rooted in a relationship between a physician and patient and is a better fit for individual AI explanations than global AI explanations. Other forms of clinical communication, especially around medical devices and pharmaceutical drugs, also offer insights about how experts have grappled with conveying complex, technical information, and we draw on examples of these as well. In contrast to clinical informed consent, U.S. law and guidelines by organizations such as the Food and Drug Administration (FDA) and Department of Health and Human Services specify rules for the labeling, package inserts, and marketing of medical devices and drugs. For the highest risk products, regulators review these communications before clearing them for the U.S. market.

The analogy between AI explanations and clinical communications in general and informed consent in particular offers some guidance about what to explain, why, when, and how, as well as an overview of potential pitfalls and ongoing debates. Though informed consent is imperfect and has garnered warranted criticism, the long-running debates within medicine and social science about informed consent offer lessons for those interested in constructing better explanations for AI systems. In the next section, we identify key insights from the example of informed consent in the setting of clinical communications that can be applied to thinking about what constitutes a good explanation for inscrutable AI systems.

Learning from Clinical Communications

In the following we discuss several lessons from comparing the discourse and history of informed consent in clinical communication with discussions of techniques and policies in explainable AI (see Table 1 for summary). The lessons are derived from extensive discussions among our interdisciplinary team, which includes clinical practitioners and medical sociologists who educate medical trainees and technical experts consulting on medical device regulation, but also computer scientists familiar with AI explainability. It is additionally informed by analysis of laws, guidelines from organizations like the AMA, and the history of informed consent in clinical settings. Through iterative discussions, including a workshop setting, we identified many parallels between explanations between physicians and their patients

[3]Between medicine, bioethics, and the social sciences, there is a robust literature on informed consent and how to improve it, which can be consulted for guidance and expectation setting by the AI/ML community, e.g., (Schenker et al. 2011; Glaser et al. 2020; Institute of Medicine, Board on Population Health and Public Health Practice, Roundtable on Health Literacy 2015; Varkey 2021; Grant 2021).

and those expected of AI systems. At the same time, we found places where the discourse diverges. Based on our analysis, we developed the following framework, composed of lessons and recommendations, that were derived from and refined through the iterative discussions. Furthermore, each lesson includes advice directed to two broad categories of potential stakeholders, computer scientists (encompassing developers and others) and policymakers (encompassing legislators, regulatory-standard setters as well as policy enforcers). While these lessons cannot address all the considerations requiring attention in the development of AI technology and policy, they serve as conceptual perspectives on the debate on the explainability of AI systems.

Lesson 1. Clinical communications are provided in recognition of patient autonomy, advancing human dignity.

Today, obtaining informed consent, as part of clinical communications, requires that physicians make an effort to meaningfully engage with patients and assess their understanding. The modern notion of informed consent stems from the longstanging obligation that physicians have to tell patients the truth based on their professional role. Truth-telling was a way to demonstrate respect for patients as people. This obligation preceded the recognition of patient autonomy as a central principle of clinical ethics by physicians in the U.S. over the course of the twentieth century. Prior to the mid-twentieth century, physicians did not always give patients a choice, or convey the risks of a course of treatment. Patients were not expected to question physicians' diagnoses or proposed treatments. Social movements, including the feminist self-help movement, initiated a trend toward patient empowerment and patients' rights, prompting physicians to look for ways to share decision-making with patients and improve clinical communication (Halpern 2004). This was an important shift in medicine, generally regarded as positive for both patients and physicians.

Physicians were given authority from the government to regulate themselves, with physicians providing oversight over other physicians (Conrad and Schneider 2009). But a series of court decisions in response to actions brought by patients against physicians in the U.S. challenged the norm of physicians' paternalism, culminating in the institution of a legal requirement of documented informed consent. The legal cases formalized changing cultural ideas about the relationship between patients and physicians, including a shift toward patient-centered care (Halpern 2004). These cases also established that disclosure of a diagnosis and treatment risks were necessary to recognize patient autonomy. Thus, the general ethical principle of patient autonomy became institutionalized in the form of a legal requirement to obtain and document the informed consent of patients.

To meet the principle of *patient autonomy* today, physicians should provide enough information about treatment (e.g., benefits and risks) for the patient to be able to deliberate and take action, including asking questions. Patients are recognized as having a right to obtain this information. This serves the larger purpose of not only promoting the autonomy of individual persons but also *"promoting autonomy as a general social value"* (JF Childress and MD Childress 2020) (p.421), requiring healthcare professionals to account for and justify their thought process. This can promote rational and mutually agreed upon decisions, given that multiple courses of action might be reasonable for a given patient. For instance, in the case of radiation therapy for lung cancer, clinicians need to manage the efficacy of treatment (delivering sufficient radiation dose to the targeted tumor) versus damage to other organs, like the heart and esophagus. Patients, especially depending on their age, will have different preferences about balancing these tradeoffs, and the informed consent process can help physicians and patients come to a specific course of action for them. At its root, informed consent establishes a baseline of information to include about an expert decision, and makes visible that there is a decision to be made. Principles of autonomy (or respect for persons, in the parlance of the Belmont Report) hold that people have a right to know information about their health and care.

It is worth noting that the ethical principle of *patient autonomy* undergirding this discussion does not always translate into more choices for patients. If a patient wants to obtain treatment, she is subject to the advice and ministrations of a physician, or a second opinion by another physician. This is likely to be the case with many AI systems, too; users' actions may be functionally constrained to using the AI system on the terms it is presented or not using the AI system at all. Even if users' actions are constrained, it is a recognition of their autonomy to give them information to deliberate. It is considered a lack of respect for an autonomous person to, in the words of the Belmont Report, "withhold information necessary to make a considered judgment, when there are no compelling reasons to do so" (Belmont Report 1979).

Ethics codes and manuals also note that physicians must balance autonomy against the other ethical principles (such as beneficence, non-maleficence, and justice) which may cut in different directions. And while these ethical principles are well-established, ethicists and physicians argue that they are not exhaustive, and suggest the elevation of other moral aims and goals, e.g., (Fiester 2007).

Advice to computer scientists: Improving explainability is often framed as an instrumental way to enable oversight or improve human-AI interaction, specifically with the goal of getting humans to invest the appropriate amount of trust in a model's prediction. There has been relatively little discussion of how explainable AI and transparent models might have an intrinsic value (but see (Colaner 2022; Selbst and Barocas 2018)). Early proposals for AI regulation do not all prescribe principles or norms behind why explanations should be offered: Among the key propositions of the White House AI Bill of Rights is the idea that users should be notified when an AI system is in use and that they should be given the chance to consent to it (U. S. White House 2022), whereas the EU GDPR refers to explanations in the context of enabling end users to contest automated decisions (Bayamlıoğlu 2022; Selbst and Barocas 2018). Including recognition of human autonomy and dignity as one poten-

tial purpose for AI explanations is worthwhile. We can learn from clinical communications that the act of explanation itself recognizes people as deliberative, autonomous agents, central to their dignity as human beings. In the case of medicine, this was a hard-won insight that required decades of patient activism and a shift in the norms and culture of clinical expertise (Halpern 2004). Rather than waiting for pushback, AI developers could offer AI explanations to end users in recognition of their dignity and autonomy.

Furthermore, a requirement for providing an explanation for an AI system could potentially affect how developers approach designing and building systems. Given two pathways that may seem comparable on other qualities, a developer might make a choice that is more easily explainable if they know they will have to account for it, especially if they know the purpose of accounting is to recognize the dignity and autonomy of end users.

Advice to policymakers: Since the time of Plato, it has been argued that in a free society, citizens do not just require an answer but the ability to ask questions and to be respected as part of a decision-making process that directly affects them (Tasioulas 2023). Requiring an explanation on this basis alone has merit. But it is also worth foregrounding *a purpose* for the explanation, alongside the requirement that one be offered. The intended purpose of an explanation will provide additional guidance to AI developers, especially in the absence of a well-established professional norm or culture establishing what should be done. The right to an explanation in the EU GDPR, while vague in its requirements, is one good example of framing an explanation around the purpose of enabling users to contest automated decisions made about them. Specifying a purpose for explanations beyond technocratic goals such as debugging/auditing, effective collaboration, or oversight, may spur more creative and ultimately useful explanations.

Lesson 2. Information provided to patients in clinical communications is targeted rather than exhaustive, to patients' ultimate benefit.

In AI, the partiality of explanations is widely recognized as a problem (Molnar 2022). The insight from clinical communications is that explanations, to some degree, need to be tailored, and that this can be a positive thing for those receiving the explanation. Multiple explanations might be necessary for different stakeholders.

The information provided during a clinical informed consent encounter is not necessarily exhaustive. Physicians are not neutrally presenting a series of options, but making a case for a given course of treatment, while acknowledging the risks of that treatment. In healthcare, patients need to know enough about a diagnosis or treatment recommendation to ask more questions–of their own physician or in a second opinion. Usually patients do not simply refuse outright, but discuss and negotiate with their physician over alternate courses of treatment. The right to question a diagnosis does not mean that a patient necessarily overrides the opinion of the physician themselves but they can opt to seek a second opinion from another provider (American Medical Association Opinion 2023a).

There are trade-offs between completeness and comprehensibility. In some cases, the precise mechanism of why or how a medication works for a specific condition in a particular patient may not be fully known; medicine has simply adopted a standard set of conventions, including randomized control trials, for measuring and showing evidence of efficacy at the population level, which is then evaluated by a regulatory body such as the U.S. FDA. For individual patients, it may not be possible to accurately convey to patients the ontological uncertainty about how medical diagnoses such cancer will exactly unfold in their specific case. Moreover, providing extensive descriptions to patients may be more than they can process. In fact, the AMA Code of Ethics (American Medical Association Opinion 2024) calls on physicians to evaluate what information patients can absorb: *"Assess the amount of information the patient is capable of receiving at a given time, and tailor disclosure to meet the patient's needs and expectations in keeping with the individual's preference."* The AMA's injunction that physicians "tailor disclosure" reflects the conclusion that total transparency would be overwhelming to patients.

The AMA Code of Ethics holds that

> *The obligation to communicate truthfully about the patient's medical condition does not mean that the physician must communicate information to the patient immediately or all at once. Information may be conveyed over time in keeping with the patient's preferences and ability to comprehend the information. Physicians should always communicate sensitively and respectfully with patients.*

This statement provides expectations for what physicians should do. But this may be received by patients in different ways. Patients might not understand jargon and may defer to physicians rather than become involved in their own care. They might ignore the information because it is too much and too difficult to absorb. Or they might ask pointed questions about the illness, or ask about the use of specific medications, echoing direct-to-consumer pharmaceutical marketing. Informed consent is intended to meet patients where they are, involving translation from technical to lay language, while providing guidance and reassurances. The flexibility of informed consent allows for physicians to adapt to the circumstances – of different patients, conditions, and organizational settings. They can provide more or less detail based on patients' stated or revealed preferences.

Patients are most closely attuned to the explanations of physicians when they get an unexpected or undesired result. When a routine mammogram finds evidence consistent with a potential breast mass, patients may ask about the false positive rate and the limits of what a mammogram can show. If the mammogram does not find anything, patients may not ask any questions. Similarly, end-users of an AI system might pay more attention to explanations when facing negative outcomes. For example, when someone is rejected from a job posting without proceeding to an interview by an AI screening system, they might want to know more about how the model is operating than if they advance and do get the

job. While people may pay more attention when they receive unwanted results, providing AI explanations routinely builds familiarity and knowledge that builds over time, so they are not starting from scratch with a negative interaction.

With AI, as in clinical settings, there are also concerns with providing too much information. Researchers have found that detailed AI model explanations overwhelm end-users with information overload and distract them from understanding how the system really works (Poursabzi-Sangdeh et al. 2021; Springer and Whittaker 2019). For example, (Ehsan et al. 2021) demonstrated that individuals, both with and without technical expertise in AI, often have inflated confidence and trust in numerical data, believing them to signify intelligence even when they cannot understand their meaning. When provided with lengthy descriptions, end users might also simply ignore them.

Advice to computer scientists: The impossibility or inadvisability of providing too much information is not a reason to provide none. Targeted explanations can be beneficial to end users and humans-in-the-loop. Failing to include an explanation of an opaque model can lead humans using the system (like judges) to form incorrect mental models of how the prediction works—assuming, for instance, that the system takes into account the severity of the crime, when in fact it does not (Rudin 2019). However, if the humans-in-the-loop knew something about the rules of the model, it eliminates at least some clearly incorrect assumptions. More detailed explanations may be necessary for other kinds of stakeholders. Figure 1 illustrates the different potential audiences to whom AI explanations might be productively targeted, highlighting the different purposes they may seek in an explanation.

In healthcare settings, the degree of information required as part of a clinical communication varies depending on the risk of harm to the patient. Higher-risk applications, especially medical devices and treatments, are accompanied by more information, which is scrutinized by an external regulator like the U.S. FDA. The same might be expected for AI. Developers working on tools where the impact on user well-being is higher (e.g., the decision can affect their lives in ways significantly more important than a simple movie recommendation) should expect greater and increasing scrutiny. This is evident in the EU AI Act, where the rules for "high-risk" AI resemble those of regulated medical devices as opposed to those of standard software tools.

Advice to policymakers: Policymakers should understand and accept that explanations are necessarily partial and should evaluate explanations within the context of their stated purpose, rather than to a standard of absolute comprehensiveness. At the same time, the amount of detail required in an explanation may need to scale with the level of the risk of the AI system.

Lesson 3. Clinical communications are governed by both norms and rules.

Everyone on our research team has filled out a multi-page consent form at the dentist without reading it and having only a cursory conversation: "We will be taking X-rays of your teeth today." This may adhere to the letter of the law on informed consent, but not to its spirit. People may go through the motions of documenting compliance with rules for explanations, but do so superficially and without an interest in ensuring that the explanations are comprehensible or fit to inform patients of the risks and benefits and treatments—it simply meets the text of the requirement. As this example illustrates, the codification of informed consent into written forms has come with pitfalls.

In practice, informed consent forms for clinical procedures may be written with legal language that seek primarily to absolve providers of legal responsibility in the event of complications or side effects. Studies in the social sciences show how this kind of institutionalization around informed consent forms preempts a more robust reckoning around ethics, conflicts of interest, or a discussion of what a given individual is likely to experience (Blee and Currier 2011). Instead of serving primarily to document the process of a physician communicating with a patient in a nuanced way about a procedure, informed consent forms can be co-opted by sponsors and organizations into becoming a waiver of legal liability by patients (Grant 2021; Hoeyer and Hogle 2014).

There is a tension between purposes, such as shared decision-making, building trust, and documenting compliance with administrative mandates, which has led to suggestions for the use of multiple different tools and forms (Hall, Prochazka, and Fink 2012). Under the weight of these competing aims, written consent forms can be lengthy and difficult to understand, including rare contingencies; or, in contravention of the idea that they facilitate conversations and recommendations for specific patients, they can be too general and not tailored enough to the specifics of a given patient (Albala, Doyle, and Appelbaum 2010; Jefford and Moore 2008; Hall, Prochazka, and Fink 2012). A common objection to compliance documentation is that documenting informed consent creates more paperwork while changing little about practice.

However, in addition to becoming formalized in consent forms, governed by rules of specific clinics, informed consent is a process governed by norms. The consent form has been produced as evidence of compliance with the requirement of informed consent. However, the broader conversation between the patient and physician can, and by convention, *should*, be broader. The ethical principle of *"respect for persons"* has become a professional *norm*, which guides physicians as to what to tell specific patients. That is, even as physicians believe there is a range of communications that could be had with a given patient about a given diagnosis, they agree on *why* patients should receive information. Physicians are assessed (by themselves, their peers, and patients) on whether they respect patients. A *norm* is a shared cultural expectation and orientation about how people should behave, and can be particularly powerful when linked to a professional role and to professional socialization. In training, physicians coach or correct trainees, providing individualized, case-by-case feedback on how trainees might more effectively communicate with patients. They evaluate each other on how well they adhere

to this norm and what efforts they make to improve (Bosk 2003). Physicians may face reputational costs if they are perceived as *insufficiently respecting patients*, above and beyond any formal sanctions brought by healthcare institutions against physicians for failing to adequately obtain or document informed consent. The norm that information should be communicated to patients allows for flexibility for physicians even as it establishes a clear professional obligation. Thus, even if the informed consent form is too long, too legalistic, and too difficult to understand, the physician has an obligation to have a conversation that conveys information in a way that the patient in front of her can absorb.

Advice to computer scientists: In addition to rules, computer scientists could explore and develop a set of norms to guide them in building AI models through ongoing and continuous education. This would help diminish the pressure on rules and guide them in uncertain or new scenarios.

As with informed consent, rules can accompany and formalize some aspects of norms. The example of informed consent is a cautionary tale in this regard. Unscrupulous physicians, or those with little power to push back on administrative directives, may focus only on the written consent form, complying with the rule and not the norm. Setting explicit rules for explanations risks turning an instrument for meaningful engagement into a compliance exercise and liability shield, where explanations satisfy the regulation but are not effective for their intended purpose. The idea that rules can become pro forma compliance exercises is certainly not new to software developers and mirrors concerns about software licenses and privacy policies. For example, there is some evidence that informing users about privacy implications of using a product may influence user behavior, with some users willing to pay more for services if accompanied by greater privacy protections (Tsai et al. 2011). Privacy policies, which outline the kinds of online data about a person and their website use that a company might collect, are now required in many contexts and intended to enable deliberate decision making. However, there is evidence that privacy policies in their current form are not effective: Typically, end users scroll through without reading them before clicking "I agree" to get on with using the product (Steinfeld 2016). Evidence suggests that users often do not read them and do not understand the information if they do try (K-Pl et al. 2007; Reidenberg et al. 2015). Privacy policies satisfy legal requirements and have crossed the consciousness of users, but seem mostly ineffective for their intended purpose.

Explanations for AI models could also end up as a formality. Developers can meet explainability requirements with inscrutable or even misleading explanations, which can be counterproductive for end-users or even actively manipulate them (Ehsan et al. 2021; Stumpf, Bussone, and O'sullivan 2016; Springer, Hollis, and Whittaker 2018). Current policy drafts or guidances for explainability and transparency in AI rarely include rules specific enough that developers could demonstrate compliance in a way that allows them, regulators, or third parties to determine whether explanations are not just present but also effective for their intended purpose (if a purpose is given at all in the first place). In a study in which policy for explanations and AI models were designed in tandem, researchers found that it is easy to find loopholes and bypass policy requirements by providing poor explanation examples that met the letter but violated the spirit of the policy (Nahar et al. 2024). But these risks do not negate the potential benefit of mandating communication. In clinical communications, the act of acknowledging a choice to be made and the risks of a treatment has benefits in and of itself (as per Lesson 1). Though some actors may violate the spirit of the rules and norms, others may make meaningful attempts at compliance, and it may be possible, with some creativity and iterative adjustments, to make the requirement more than a pro forma one.

The tradeoff between norms and rules is a common tension in regulation, and we do not purport to have solved this problem. There is an important difference between the clinical and AI cases, however. Policies for physicians grant them considerable deference as professionals, leaning on an existing infrastructure of self-regulation. External sanctions are typically only invoked for high level violations of rules, like the cutting off of federal funds for care or research to non-compliant institutions. There are no equivalent mechanisms for normative or self-regulation of computer scientists. Norms for computer scientists would have to be created, debated, and sustained through educational and work institutions.

Advice to policymakers: Regulation of informed consent provides both inspiration and can serve as a warning for how to regulate AI explainability: Lower-level social enforcement mechanisms through norms and continuous education might be more effective for adoption of good practices than strict requirements imposed from above. On the other hand, stricter requirements such as external audits of products and processes as well as user studies of whether explanations are effective for their stated purpose can set higher expectations. It may be worthwhile to have external bodies ask for evidence, similar to the model of the U.S. FDA in auditing or inspecting medical device and drug manufacturers, and to show evidence specifically that users comprehend explanations, a requirement that informed consent rules have largely informal mechanisms to enforce.

Lesson 4. What constitutes a "good" consent procedure and a "good" explanation will change over time.

Though informed consent is rooted in longer-lasting ethical principles, in practice and in documentation, it has undergone many changes over the last hundred years. Clinical informed consent is a policy solution consistent with a principles-based form of regulation, in which regulated parties are asked by regulators to show adherence to general principles rather than adhere to prescriptive, specific rules (Black 2010). In the case of healthcare, the principles have remained somewhat consistent over time (e.g., respect for persons, beneficence, non-maleficence, and distributive justice) even as how these principles are weighed relative to one another and instantiated into healthcare policy has

changed over time. But some trial and error was necessary before norms and conventions coalesced. And societal norms around acceptable or desirable physician-patient relationships changed, leading to shifting understandings of how physicians and patients should interact over time.

After agreeing on the idea of informed consent in the 1950s and 1960s, physicians and ethicists continued to debate what counted as "informed." Initially, courts focused on whether or not physicians disclosed information to patients and on what constituted professional standard practice for disclosure rather than on patients' understanding of information. Critics charged that informed consent took the onus off of physicians to explain and guide patients, and left the burden of parsing information on informed consent forms too much on patients. By the 1980s, in response, and after pushback by patients, shared decision-making between physicians and patients emerged as the new goal, with more emphasis placed on conversations between the two (Katz 2002; Presidential Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research 1982). Even within shared decision-making, there are debates about approaches that are more and less individualistic (Entwistle and Watt 2016; JF Childress and MD Childress 2020).

There has also been evolution in how informed consent is documented. Starting in the 1990s, physicians generally agreed that consent forms should 1) contain enough information for a patient to make a deliberative decision, 2) be written in plain language, and 3) be short (Schenker et al. 2011). However, the content of what should go into a form has continued to be debated, with some coming to the conclusion that it is impossible to meet all three of these criteria (Grant 2021). Recently, physicians and researchers have complemented the written consent form with other interactive means of communication, from short videos to visuals, finding that those with test or feedback components to be an improvement (Glaser et al. 2020; Grady et al. 2017). These innovations help re-engage patients who have come to ignore paper forms.

Advice for computer scientists: Developers should prepare to update their AI systems to consider new explanations as technologies and expectations evolve. In a typical "design for change" matter, explanations can be isolated as a separate concern that may be updated without changing the rest of the system. As developers incorporate new AI technology, they should also keep up with recent developments in explainability research and practices. Developers would benefit from concrete guidance to set expectations and specific forms of evidence on how they could demonstrate achieving the policy goals. Regular updates and audits might also be indicated, as explainability norms still change frequently.

While techniques may change, the core values of what an explanation is for will remain. Developers should focus on the overall mission and goal of providing an explanation and accept that though the tools and best practices to achieve that will change rapidly, the core values remain.

Advice to policymakers: AI is quickly changing, but regulation need not anticipate every twist and turn in the technology. The challenge is that because AI is relatively new and unfamiliar to the general public, norms and expectations around it are still emergent and unsettled, making it even more important that rules for AI have a plan for revision. The framers of rules for clinical communication in biomedical research did not get it perfectly right the first time they wrote rules, and they remained in operation for forty years (Stark 2023). Building in a more regular revision schedule gives more opportunities for dynamic tuning of rules to emergent conditions, and puts less pressure on the first draft. For instance, the National Institute of Standards and Technology (NIST), directed by the U.S. Congress in 2021 to develop voluntary frameworks and best practices to guide the progress and use of trustworthy AI systems, released its first set of guidelines for development in January 2023, through its "AI Risk Management Framework 1.0." NIST's embrace of policy versioning recognizes the ongoing, unpredictable development that marks the future of AI, and allows for greater agility in reviewing standards and policy (Nelson 2024). To meet evolving social norms about disclosure and the obligations that companies have to their customers, guidelines around explanations for AI will need to be updated, ideally at set intervals and ideally by a third party standard-setting or government organization (Stark 2023; Nelson 2024).

While the specific details of what is required is bound to change over time, the reason for the explanation remains the same: to allow a human to make an informed decision about the risks they face as a result of some action. Hence, regulation may be best tuned to require disclosure of such risks and potential mitigation strategies as opposed to specific details of a particular model or task. Existing governance structures may not adequately be prepared for the challenges posed by AI systems, and policymakers should be ready to try new strategies for AI governance (Nelson 2024).

Lesson 5. Negative experiences with clinical communications can have a broader effect on how patients think about medicine and medical technologies.

One critique of clinical communication, and the informed consent process, is that it is focused only on patients and their physicians (Varkey 2021). But there is a broader social effect for even these seemingly dyadic communications (Entwistle and Watt 2016). Explanations to patients also build trust by generalizing the goodwill and authority invested in some external authority to a given person. Informed consent can, but does not necessarily, build a patient's trust in the competence of their individual care provider. If the informed consent discussion is a substantive exchange, then it can build rapport between a patient and a physician. And if it goes exceptionally badly, patients may begin questioning the competence of their physician. Patients are not using the informed consent conversation as the primary way to evaluate their physicians; they leave it to hospitals and clinics to hire competent physicians. But that informed consent is performed in a standardized, regular way in a healthcare facility is a mark of its quality and modernity (Varkey 2021).

It is a set of conventions that makes legible the expectations of conduct for physicians and patients.

In the best case scenario, informed consent establishes what might go wrong ahead of time, giving the patient information about expected and common side effects or adverse events. This helps patients see or understand the potential outcomes of specific treatment decisions, but also acknowledges the inherent uncertainty of treatments. If something does go wrong, physicians are obliged to disclose errors to individuals after they occur (Fiester 2007). Giving an account, or narrating what has happened to an affected individual, is a basic form of accountability (Neyland 2019). Because disclosure of information about diagnosis and treatment is routinized and happens in every case, not simply done in the aftermath of harm or error, it helps build patients' trust overall in the system—that even if a specific course of treatment fails, the system of medicine can recognize that failure, a necessary first step to remedy.

There are not currently routine practices around AI explanations or disclosures. Because public norms and expectations for AI are still emerging and unsettled (Nelson et al. 2020), people may not have a sense that AI systems are fallible, can make errors, or be used improperly. Every explanation for every AI system, therefore, works toward building users' knowledge about AI over time. Creating a norm around routinely providing explanations for AI systems can have an effect on how people perceive AI more broadly.

The public remembers when something goes wrong, such as when there is a privacy scandal, or when companies are revealed to have violated laws or norms. Regulation is often reactive to these high-profile controversies (though policy is also patiently and incrementally built by other, slower means) (Junginger 2013; Maor 2014). While norms in clinical communications such as informed consent build in considerable flexibility at the level of dyads and individuals, they have had a much broader social effect, helping to undergird people's trust – in their physicians, in a specific treatment, in medicine, and in science.

Advice to computer scientists: In the absence of conventions for explanations, especially for low risk and low stakes AI systems that are not subject to regulatory scrutiny currently, explanations may only be offered when things go wrong, like the demonstration of an error or bias in the AI system's predictions. If disclosure and explanations are only forthcoming after an AI system has seemingly failed, it diminishes users' trust in the specific AI at hand, but also, potentially, in AI in general. It is worth providing AI explanations more routinely.

Hence, AI explanations may cumulatively, in conjunction with many repeated exposures to different kinds of AI, shape how people think about AI. The takeaway is not that explanations should be provided to generalize trust in all AI. Rather, failing to provide explanations, providing explanations only after something bad happens, or providing maliciously bad explanations could undermine public trust in AI. More concretely, providing technical AI explanations that have no meaning to users is equivalent to a physician speaking only in jargon. Meaningful AI explanations can help establish that the system is or is not fit for a given task, a good outcome for AI. This can build trust in a specific system or in all the systems of a specific company. Big tech companies may offer explanations about their products in order to develop ongoing relationships with consumers. But the entire AI field could lose more credibility by providing bad explanations for AI (e.g., incomprehensible, lacking grounding) than by not providing any.

Advice to policymakers: Explanations should be required for both positive and negative outcomes to establish norms in this nascent space. Establishing a set of rules and expectations for AI can help. But foregrounding and accounting for how those rules are made will also be important. Regulatory agencies are also critical to building and maintaining public trust in complex technologies. Policymakers should try to produce accessible communications/reports geared to the large public where they explain how the technology is regulated, and what the steps used are to ensure compliance. Trust in regulators will also be important.

Conclusions

While AI systems pose some new societal and regulatory challenges, the challenge of explaining complex, even unknowable, processes has some precedent–in healthcare, among other areas. In this position paper, we highlight how the example of clinical communications anticipates debates about transparency requirements for AI systems, as well as illustrating potential consequences of different courses of actions. Rather than holding up clinical communications, and the more specific example of informed consent, as models to be emulated, we analyze them as useful tools to think with at a time when regulations for explainability are just beginning to emerge. While informed consent is a settled precedent, codified into rules and enshrined in the professional norms of physicians, debates remain about the best way to convey information to patients and how to assess what ultimately constitutes "informed" consent on the part of patients. This suggests that the task of figuring out "good" explanations for AI systems will likewise be an evolving social and technical process.

We identify five key insights that computer scientists and regulators can take from the example of clinical communications that: 1) offering explanations for AI systems and disclosure of their use recognizes the dignity of those using and impacted by it; 2) AI explanations can be productively targeted rather than singular and comprehensive; 3) AI explanations can be enforced through rules but also norms, guided by underlying core values; 4) what constitutes a "good" AI explanation will likely not be fixed on the first try but will required repeated adjustment; and 5) AI explanations will shape broader perceptions of the value, utility, and harms of AI. Drawing on our interdisciplinary expertise, we direct regulators and computer scientists to the literature in the social sciences, bioethics, and medicine on clinical communications for further potential insights for AI explainability.

References

Albala, I.; Doyle, M.; and Appelbaum, P. S. 2010. The evolution of consent forms for research: a quarter century of changes. *IRB*, 32: 7–11.

American Medical Association Opinion. 2023a. 1.1.3 Patient Rights, AMA Code of Medical Ethics. https://code-medical-ethics.ama-assn.org/ethics-opinions/patient-rights. Accessed 13 Sep 2023.

American Medical Association Opinion. 2023b. 2.1.1 Informed Consent, American Medical Association Code of Ethics. https://code-medical-ethics.ama-assn.org/ethics-opinions/informed-consent. Accessed 13 Sep 2023.

American Medical Association Opinion. 2024. 2.1.3 Withholding Information from Patients, American Medical Association Code of Ethics. https://code-medical-ethics.ama-assn.org/ethics-opinions/withholding-information-patients. Accessed 19 Jan 2024.

Bayamlıoğlu, E. 2022. The right to contest automated decisions under the General Data Protection Regulation : Beyond the so-called "right to explanation.". *Regul Gov*, 16: 1058–1078.

Bazzano, L. A.; Durant, J.; and Brantley, P. R. 2021. A Modern History of Informed Consent and the Role of Key Information. *Ochsner J*, 21: 81–85.

Beauchamp, T. L. 2011. Informed Consent: Its History, Meaning, and Present Challenges. *Camb Q Healthc Ethics*, 20: 515–523.

Bell, A.; Solano-Kamaiko, I.; Nov, O.; and Stoyanovich, J. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266. New York, NY, USA: Association for Computing Machinery.

Belmont Report. 1979. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Fed Regist*, 44: 23191–23197.

Benjamin, R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Bhatt, U.; Xiang, A.; Sharma, S.; et al. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. New York, NY, USA: Association for Computing Machinery.

Black, J. 2010. The Rise, Fall and Fate of Principles Based Regulation. Working Paper 17, LSE Legal Studies.

Blee, K. M.; and Currier, A. 2011. Ethics beyond the IRB: An introductory essay. *Qual Sociol*, 34: 401–413.

Bosk, C. L. 2003. *Forgive and remember: managing medical failure*. University of Chicago Press.

Broussard, M. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press.

Carlile, P. R. 2002. A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development. *Organization Science*, 13: 442–455.

Carlile, P. R. 2004. Transferring, Translating, and Transforming: An Integrative Framework for Managing Knowledge Across Boundaries. *Organization Science*, 15: 555–568.

Colaner, N. 2022. Is explainable artificial intelligence intrinsically valuable? *AI Soc*, 37: 231–238.

Conrad, P.; and Schneider, J. W. 2009. Professionalization, monopoly, and the structure of medical practice. In *The sociology of health and illness: Critical perspectives*, 194–200.

Dedikov, G. 2023. *Explainable AI and User Experience. Prototyping and Evaluating an UX-Optimized XAI Interface in Computer Vision*. GRIN Verlag.

Ehsan, U.; Passi, S.; Liao, Q. V.; et al. 2021. *The who in Explainable AI: How AI background shapes perceptions of AI explanations. arXiv [cs*. HC].

Ehsan, U.; and Riedl, M. O. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, 449–466. Springer International Publishing.

Entwistle, V. A.; and Watt, I. S. 2016. Broad versus narrow shared decision making: patients' involvement in real world contexts. In Thompson, E. G. E. A., ed., *Shared Decision Making in Health Care: Achieving Evidence-Based Patient Choice*, 7–12. New York, NY: Oxford University Press.

European Parliament. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. *1–144*.

Executive Order 14110. 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *Federal Register November*, 1: 75191–75226.

Fiester, A. 2007. Viewpoint: why the clinical ethics we teach fails patients. *Acad Med*, 82: 684–689.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; et al. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. IEEE.

Glaser, J.; Nouri, S.; Fernandez, A.; et al. 2020. Interventions to Improve Patient Comprehension in Informed Consent for Medical and Surgical Procedures: An Updated Systematic Review. *Med Decis Making*, 40: 119–143.

Grady, C.; Cummings, S. R.; Rowbotham, M. C.; et al. 2017. Informed Consent. *N Engl J Med*, 376: 856–867.

Grant, S. C. 2021. Informed Consent—We Can and Should Do Better. *JAMA Netw Open*, 4: e2110848–e2110848.

Hall, D. E.; Prochazka, A. V.; and Fink, A. S. 2012. Informed consent for clinical treatment. *CMAJ*, 184: 533–540.

Halpern, S. A. 2004. Medical Authority and the Culture of Rights. *J Health Polit Policy Law*, 29: 835–852.

Hoeyer, K.; and Hogle, L. F. 2014. Informed Consent: The Politics of Intent and Practice in Medical Research Ethics. *Annu Rev Anthropol*, 43: 347–362.

Institute of Medicine, Board on Population Health and Public Health Practice, Roundtable on Health Literacy. 2015. *Informed Consent and Health Literacy: Workshop Summary. National Academies Press*. National Academies Press.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. New York, NY, USA: Association for Computing Machinery.

Jefford, M.; and Moore, R. 2008. Improvement of informed consent and the quality of consent documents. *Lancet Oncol*, 9: 485–493.

JF Childress and MD Childress. 2020. What Does the Evolution From Informed Consent to Shared Decision Making Teach Us About Authority in Health Care? *AMA Journal of Ethics*, 22: E423–429.

Jia, Y.; McDermid, J.; Hughes, N.; et al. 2023. The Need for the Human-Centred Explanation for ML-based Clinical Decision Support Systems. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 446–452. IEEE.

Joshi, S.; Agarwal, C.; and Lakkaraju, H. 2021. Tutorial: Explainable ML in the Wild: When Not to Trust Your Explanations. In *Accepted Tutorials of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM.

Junginger, S. 2013. Design and Innovation in the Public Sector: Matters of Design in Policy-Making and Policy Implementation. *Policy Design Annual*, 1: 1–11.

K-Pl, V.; Chambers, V.; Garcia, F. P.; et al. 2007. How Users Read and Comprehend Privacy Policies. In Interface, H.; and the, eds., *Management of Information. Interacting in Information Environments. Berlin*, 802–811. Heidelberg: Springer.

Katz, J. 2002. *The Silent World of Doctor and Patient*. Johns Hopkins University Press.

Kaur, H.; Nori, H.; Jenkins, S.; et al. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. New York, NY, USA: Association for Computing Machinery.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; et al. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*, 117: 12592–12594.

Latif, E.; Zhai, X.; and Liu, L. 2023. *AI Gender Bias, Disparities, and Fairness: Does Training Data Matter? arXiv [cs*. CY].

Luria, M. 2023. Co-Design Perspectives on Algorithm Transparency Reporting: Guidelines and Prototypes. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1076–1087. New York, NY, USA: Association for Computing Machinery.

Maor, M. 2014. Policy bubbles: Policy overreaction and positive feedback. *Governance*, 27: 469–487.

Molnar, C. 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

Nahar, N.; Rowlett, J.; Bray, M.; et al. 2024. Regulating explainability in machine learning applications – observations from a policy design experiment. ACM Conference on Fairness, Accountability, and Transparency 1.

National Institute of Standards and Technology. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. U. S. Department of Commerce.

Nelson, A. 2024. The right way to regulate AI. *Foreign Aff January*, 12: 1–11.

Nelson, C. A.; Pérez-Chada, L. M.; Creadore, A.; et al. 2020. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol*, 156: 501–512.

Neyland, D. 2019. *The Everyday Life of an Algorithm*. Springer International Publishing.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Panigutti, C.; Beretta, A.; Fadda, D.; et al. 2023. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Trans Interact Intell Syst*, 13: 1–35.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; et al. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. New York, NY, USA: Association for Computing Machinery.

Presidential Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. 1982. Making Healthcare Decisions: A Report on the Ethical and Legal Implications of Informed Consent in the Patient-Practitioner Relationship.

Reidenberg, J. R.; Breaux, T.; Cranor, L. F.; et al. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technol Law J*, 30: 39–68.

Ribera, M.; and Lapedriza García, A. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *CEUR Workshop Proc*.

Rong, Y.; Leemann, T.; Nguyen, T.-T.; et al. 2022. *Towards Human-centered Explainable AI: User Studies for Model Explanations. arXiv [cs*. AI].

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intell*, 1: 206–215.

Schenker, Y.; Fernandez, A.; Sudore, R.; and Schillinger, D. 2011. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review. *Med Decis Making*, 31: 151–173.

Selbst, A. D.; and Barocas, S. 2018. The intuitive appeal of explainable machines. *Fordham Law Rev*, 87: 1085.

Springer, A.; Hollis, V.; and Whittaker, S. 2018. *Dice in the black box: User experiences with an inscrutable algorithm. arXiv [cs*. HC].

Springer, A.; and Whittaker, S. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 107–120. New York, NY, USA: Association for Computing Machinery.

Star, S. L.; and Griesemer, J. R. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Soc Stud Sci*, 19: 387–420.

Stark, L. 2023. Medicine's Lessons for AI Regulation. *N Engl J Med*, 389: 2213–2215.

Steinfeld, N. 2016. I agree to the terms and conditions. *(How) do users read privacy policies online? An eye-tracking experiment. Comput Human Behav*, 55: 992–1000.

Stumpf, S.; Bussone, A.; and O'sullivan, D. 2016. Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

Tasioulas, J. 2023. *The Rule of Algorithm and the Rule of Law. Vienna Lectures on Legal Philosophy*. Vienna Lectures on Legal Philosophy.

Tsai, J. Y.; Egelman, S.; Cranor, L.; and Acquisti, A. 2011. The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. *Information Systems Research*, 22: 254–268.

U. S. White House. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.

Varkey, B. 2021. Principles of Clinical Ethics and Their Application to Practice. *Med Princ Pract*, 30: 17–28.

Vera Liao, Q.; and Varshney, K. R. 2021. *Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv [cs*. AI].

Yildirim, N.; Pushkarna, M.; Goyal, N.; et al. 2023. Investigating how practitioners use human-AI guidelines: A case study on the people + AI guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.