Transparent Artificial Intelligence: Technical Explainability and Its Troubles

Abstract

Transparency is the most commonly listed desideratum for Ethical Artificial Intelligence. However, analysis of the literature reveals the lack of a cohesive definition of Transparency. This paper points out that AI Transparency delineates into two categories: technical and social. Technical Transparency is the ability to view the causal structure of an AI; it is a more academic solution, solving the black-box problem of many novel AI implementations. Social Transparency is the disclosure of aspects such as how a system is used, by whom, for what, when, its pros and cons, dangers and risks, the level of technical explainability available, privacy concerns, fairness, and more. Researchers and regulators often deem both sides of this delineation desirable. However, this paper argues that some aims of AI Ethics such as trust, risk mitigation, control, utility, and regulation are better served by promoting Social Transparency instead of Technical Transparency. AI acts as a kind of social agent, thus making its social standing a more important consideration than causal understanding—a feature that mirrors Transparency in human activity. Further, Technical Transparency carries several risks, such as the creation of false knowledge, calcification of prejudice, justification for social opaqueness, and decreased security when not accompanied by Social Transparency. We should take extra care to ensure that the Ethical-AI principles we enumerate are effective and reduce risk. In both cases, technical explainability appears to fall short, being less effective and potentially dangerous.

Justin Edward Oxford
2 January 2022

Introduction

A common project in Artificial Intelligence (AI) Safety and Ethics is the enumeration of a list of principles. The idea is that AI which adheres to these ideals will be sufficiently safe and ethically sound. These principles are often virtues such as fairness, security, responsibility, privacy, and trust. However, no principle is as commonly listed as transparency, appearing in approximately seven out of eight such lists (Jobin 2019). Despite its popularity, transparency is perhaps the least well-defined. It acts as a blanket term for ideas such as explainability, understandability, openness, and disclosure. As this paper points out, these conflated ideas appear to delineate broadly into two categories: technical explainability and social openness. Unfortunately, the literature does not appear careful to maintain this distinction, potentially leading to trouble for AI Ethics.

Artificial Intelligence is unique in that it can undertake cognitive-social tasks typically done by humans (Bostrom and Yudkowsky 2014)—e.g., credit assessments, content moderation, medical care, and more. Consequently, we can understand AI Transparency similarly to the transparency of other social agents or institutions. We want openness to the socially relevant aspects: where, why, by-whom, for-whom. At the same time, there is a desire to know how and why AI comes to its decisions (i.e., its technical explainability, logical or causal structure). On the surface, social and technical transparency appear equally desirable. However, this paper argues that aims of AI Ethics such as trust, risk mitigation, control, utility, and regulation are better served by promoting social openness rather than technical explainability. Further, technical explainability carries several risks, such as creating false knowledge and justification for social opaqueness. Thus, those engaged in projects such as the development of ethical AI principles

should take greater care in defining aspects of transparency and strongly reconsider listing technical explainability as a desideratum until these troubles are addressed.

Artificial Social Intelligence

A feature of Artificial Intelligence that ethically distinguishes it from other technologies is its ability to do cognitive work in social contexts (Bostrom and Yudkowsky 2014). In addition to design or mechanical work, AI can be used to handle complex social tasks such as assessing credit applications, curating social network feeds, providing medical care, counseling, therapy, driving auto-mobiles[1], marketing, policing, making policy, handling customer service, hiring, and more (Stone et al. 2016). While software and algorithms are already used in most, if not all, of these domains, the agents ultimately making the decisions are still mostly humans; but increasingly, AI is assuming the role of decision-maker and removing the human altogether (Sundar 2020; *ibid*. 2016). When such responsibility is handed off, we expect that the AI will also take on the social expectations to which humans were held (*ibid*. 2014). If we consider such AI systems to be agents similar to their human counterparts, it follows that they might have moral agency and responsibility for actions, especially in cases of 'technological mishap' (Johnson & Verdicchio 2019). Further, in the long-term, if some AIs begin to possess certain mental qualities, such as the capacity for qualia, we may be forced to view them not just as responsible agents but as beings with moral status (*ibid*. 2014). This highlights a distinction between AI and other technologies. AI does not just come with technical challenges and passive social impact. It is actively and intentionally interacting with social structures.

[1] Social insofar as the software makes potentially social or moral decisions, not the physical act of driving itself.

When considering this active and intentional interaction, it is tempting to think that only AI which is sufficiently advanced will be able to meet this criterion of being a social actor. But this is misguided. Limiting ourselves to highly advanced systems risks failing to account for systems that are equally socially impactful yet less advanced. Even less advanced systems can be independent social actors when we accept a broader definition of AI. While there are many proposed definitions of AI (Bringsjord and Govindarajulu 2020; Monett et al. 2020; Stone et al. 2016), we can use Stone et al.'s interpretation of intelligence by degree as a framework to construct an understanding of AI as socially active by degree. Just as a calculator or neural network has a degree of intelligence, various AIs can be understood to have degrees of social activeness. A basic program with preset responses to therapy patients is still having a degree of social interaction; a decision tree for culling social media posts is still moderating social interaction. Understanding AIs sociality by degree helps avoid fixating only on advanced systems, such as neural networks, and maintains a broad view of the AI landscape. Further, it is plausible that socially directed treatments for one implementation might be easily transferred to other systems of differing degrees.

While all technology has some degree of social impact, the key difference of AI is its activeness and intent when placed in social contexts. Further, this activeness is not limited to advanced or near-future systems; it applies to a wide range of applications. This is not to say that all AIs are social actors, however. Many implementations can exist which are socially agnostic, but those implementations carry little to no ethical weight. We are not building AI ethical rulebooks to be applied to art creation algorithms. We are building AI ethical rulebooks primarily to mitigate risk and harm of AI in high-stakes situations and to promote social good (Jobin et al., 2016). While this discussion might benefit from further development of this point, it is not the

focus of this paper. The point has been to highlight AI's sociality and trajectory into moral status. This strengthens the case for steering away from technical explainability and into a social openness.

Two Notions of Transparency

Transparency is the prevailing principle of AI ethics. According to the analysis by Jobin et al. (2019), they found that 73 of the 84 lists of principles surveyed contain the following terms: transparency, explainability, explicability, understandability, interpretability, communication, and disclosure. Such prevalence of these terms implies a strong desire by an array of stakeholders[2] to know, see, or understand what is going on with an AI implementation. But what exactly they want to see and what they mean by 'what is going on' are anything but clear. According to Jobin et al., thematic analysis reveals "Significant variation in relation to the interpretation, justification, domain of application, and mode of achievement." Transparency is poorly defined. However, a delineation appears to emerge in the various definitions upon closer inspection. On the one hand, there are calls for greater technical explainability—i.e., causal understanding of the software—and on the other, there are desires for disclosure and openness—i.e., open-source or notification of use (for a similar observation of this delineation, see Larsson & Heintz 2020).

The literature at large is not careful to maintain much, if any, distinction of these dimensions. Some authors focus on technical aspects, some focus on disclosure, others conflate both into a broad call for transparency and address them without making a distinction, or they focus on aspects external to the AI itself like data transparency (Jobin et al. 2019; Sambasivan et

[2] Jobin et al. analyzed lists of principles from tech industry, governments, private research firms, and others.

al. 2021). But explainability and understandability imply knowledge of *how* the technology works, and communication and disclosure imply knowledge *that* or *why* the technology is being used. These are very different concepts, and we should make the distinction between these two dimensions clear because the efficacy of our ethical treatments depends on it. In the remainder of this section, this paper will attempt to outline these two senses of transparency such that they can be evaluated. Implicit in this attempt is the assertion that this distinction indeed does exist, but once the nature of these two notions is made clear, we should be free of any doubt in their existence. For clarity, this paper refers to the two notions of transparency in AI as Technical Transparency and Social Transparency.

Technical Transparency is the ability to see and understand the inner-workings of AI systems. It makes it possible for an inspector to directly observe what is happening and understand why a system generates a certain output from certain inputs—that is, it allows one to see a system's causal structure. Importantly, an inspector is not impaired by any technical opaqueness, commonly known as the 'black box'. The end of Technical Transparency is the solution to the black box problem. Technical Transparency says nothing about whether a non-expert could easily understand the outputs, that such outputs are useful, or that anyone who should have visibility has access; it merely means that some human, somewhere, has full visibility into the causal structure of the entire system. There are efforts underway to produce models which are technically explainable. Yet despite the scale of interest in this problem for Computer Science (Wachter et al. 2018), Technical Transparency is a narrow problem in terms of the AI landscape.

When considering the broad range of AI architectures, we get a sense of the narrow (and ever-shrinking) scope of calls for Technical Transparency. Proponents of Technical Transparency

are concerned primarily with what remains on the technological fringe. The project of those pursuing Technical Transparency is finding the causal structures of neural networks, genetic algorithms, and others that give rise to black boxes. Yet this refers to only a small portion of the total picture. There are many types of AI that can be fully inspected, such as hard-coded algorithms, calculators, and Bayesian trees (Bostrom and Yudkowsky 2014). Consequently, Technical Transparency fails to address these and other implementations—a significant portion of AI. Further, as the research matures and algorithms that are currently opaque gradually become more inspectable, calls for Technical Transparency will come to ignore a greater portion of use cases.[3] The good news is that this concludes in fully explainable AI; the bad news is that this carries several key risks, which we will soon explore.

Social Transparency is the ability for stakeholders to access the information regarding the state of affairs of a system, its effects and implications. A catch-all, it is the ability for one to know aspects such as how the system is used, by whom, for what, when, its pros and cons, its dangers and risks, its level of Technical Transparency, privacy concerns, its fairness, and more. Proponents of Social Transparency are concerned primarily with disclosure and ease of access to information. An AI may be perfectly technically explainable but be hidden on a computer in a data center, endlessly consuming and interpreting data; such a system is socially opaque. As such, Social Transparency is not regulated by the current state of research or technical limitations. One can question the social aspects of even the most unexplainable algorithms. It is instead regulated by (or limited by) social structures such as laws, ethical guidelines, public education, disclosure, proprietary research and development, and more.

[3] Barring novel unexplainable architectures, but even then, this process of narrowing would continue ad infinitum.

As a system becomes technically transparent, that system lends itself into further Social Transparency. This is evidenced by the very nature of Technical Transparency being the ability to give an explanation of what is occurring. Such an explanation can be passed on or made available to stakeholders. Further, once we ask who the stakeholders are, how much they should know about the emerging technical aspects, and what these technical aspects mean beyond the laboratory, we are having a conversation about Social Transparency.

Due to the fact that Technical Transparency adds to Social Transparency, it is plausible one might infer that Social Transparency depends on, or follows from, Technical Transparency. However, this view is problematic as it can lead us to believe that we must attain Technical Transparency prior to any kind of disclosure. We can have Social Transparency prior to technical understanding. There is plenty to be discussed in addition to the details of some system's inner workings. In fact, informing stakeholders that we do not know the first thing about how a system works is itself an act of being socially transparent. We can be open in regards to the social aspects of a black-box AI by informing stakeholders that the AI is being used and of the risks and reliability using metrics, all while having no clue why the system outputs what it does.

Thus we have a view of transparency that delineates it into technical explainability and social openness. Due to the breadth of Social Transparency, there is no notion of transparency that does not fall into one of these two categories. Either the desire for transparency is aimed at the code or at its implementation. Further, there do not appear to be any other major categories of transparency. Explainability and understandability are within Technical Transparency, and communication and disclosure are addressed within Social Transparency. Regardless of further categorizations, the modus operandi of ethicists and those concerned with safety seems to be requesting every kind of transparency for AI under the assumption that all transparency is

equally good. In what follows, I aim to provide a case against this norm, first proposing that Technical Transparency is the less potent of the two notions and thus unworthy of heavy resource allocation, second highlighting several dangers which plague Technical Transparency.

Technical vs. Social Transparency

Some aims of AI Ethics are to minimize risk, avoid harm to humans, and increase controllability (Müller 2021; Bostrom and Yudkowsky 2014). The goal of AI Transparency is often argued to be fostering trust and safety (Jobin et al. 2019). But Technical Transparency is far from necessary to attain any of these goals (Rudin and Radin 2019). Focusing instead on the development and promotion of Social Transparency appears to be a much more fruitful exercise as it can achieve these goals without needing to wait for the science to catch up. Although not an exhaustive list of the goals of AI Ethics and Transparency, we can touch upon those which are often central to the discussion: trust, risk of harm to humans, control, the usefulness of implementations, regulation, and long-term solutions.

***Trust**:* Technical Transparency can undoubtedly be helpful in building trust, but it is not the only consideration. Trust can be fostered via other means such as data openness, user education, reduction bias, controllability, security, demonstrated fairness, and disclosure of situational risk (Asan et al. 2020) — indeed, all factors of Social Transparency. Nonetheless, Technical Transparency is a way of validating systems to be private, secure, and fair, and it can inform us on issues of responsibility and explain mishaps. Without Technical Transparency, we are only observing the effects (or outputs) of systems which we can then judge to be private, secure, fair, and responsible, but we can say no more than a probabilistic prediction for the next output before it arrives.

It would seem that without Technical Transparency, we have nothing more than black-boxes, estimations, and hope. It makes sense that we would want the ability to know with certainty how a system will behave, but this is not a panacea. Understanding exactly how a system works is of little use for building trust if the aspects of the system are behind a veil to stakeholders. Technical Transparency is only useful in building trust if there is Social Transparency distributing the facts. Another way to say this is that Social Transparency comes first in building trust with AI systems. While not a reason to abandon Technical Transparency, effort would be best spent in making AI open and disclosed to build trust around AI as a whole and not just a single deployment's accuracy.

Risk & Harm to Humans*:* Artificial Intelligence carries fears of risk and harm to humans, especially in the case of general AI or superintelligent AI. For example, consider an AI that has just become generally intelligent (an AGI), and assume that it has been given some authority over human affairs.[4] Maybe the AGI will be kind, but maybe it will decide the best way to end human suffering is to end human existence. Without Technical Transparency, we would have no hope of knowing how it works (i.e., it would be a black box). Is the AI making purely utilitarian decisions? What exactly is it taking into account? Without this information, we would likely fail to stop it. This is an immense risk (Bostrom and Yudkowsky 2014). However, this is a red herring. If proper checks were put in place regarding Social Transparency, it is doubtful that such a system would be deployed except by an act of terrorism, incompetence, or some other edge case. Even if a bad actor deployed a malevolent AI, Technical Transparency might not be useful if the information about the system is hidden from those trying to stop it. Technical

[4] This could be instating and enforcing legislation, controlling some massive institutions such as banks or military forces, or something else equally impactful.

Transparency only works in stopping such an AI if those attempting to stop it also have access to the source code. And having access to source code is a feature of Social Transparency.

Returning to the case of non-malevolent deployment, Social Transparency is a much more useful safeguard against the threat of dangerous AI (and AGI) than Technical Transparency. Although it is not clear that either of these dimensions of transparency will be useful in curtailing risk from superintelligence (Bostrom 2016), Social Transparency would provide the best chance at catching the genie before it escapes the bottle. Social Transparency is prescriptive by nature, providing the ability for multiple eyes to see potential problems in context before they arise and to help deter bad consequences. Technical Transparency is reactive, only giving us tools to see inside something that is already created. A system could be entirely technically explainable but be so hidden that no one is able to stop it from wreaking havoc.

Controllability*:* One way of defining Controllability is: creating AI such that it does what we want (Yampolskiy 2020). But what do we want? As technologies evolve and change, so too do our expectations and uses for it, and it is useful to have higher level guiding principles that need not change as often. Low-level solutions such as those offered by Technical Transparency are very narrow in scope. Solving controllability for recurrent neural networks plausibly leaves the rest of the AI landscape largely unaddressed. These low-level solutions are rigid to their respective application and probably fail to meaningfully address technologies that do not yet exist.

Social Transparency, in contrast, is far more plastic than Technical Transparency. This is particularly beneficial in that it does not require laws to be entirely rewritten or amended in granular detail because of a new system architecture. Additionally, if the end of Technical Transparency is attained (i.e., no more technically opaque systems exist), the ethical

conversation will continue in Social Transparency and will evolve with the society's needs and tastes (Rességuier and Rodrigues 2020). Social Transparency will outlive Technical Transparency if we assume any positive rate of development, and therefore we should be focusing more energy on building a solid foundation for it.

Utility: The utility of both Technical and Social Transparency depends in part on the use case of a system, increasing as the stakes increase. But the usefulness of Social Transparency is preserved in low-stakes cases, whereas Technical Transparency becomes nearly useless. For example, it is not terribly important to know exactly how a neural network decides which recipes to suggest for your dinner; however, it does seem incredibly important to know the causal basis of a decision for a medical diagnosis (especially if such information is legally required). In low-stakes scenarios like the former example of a recipe recommender, aspects such as why AI is being used, what other processes it may be completing behind the scenes, and where it is sending your data are more of a concern than exactly how the tensors led it to recommend kung pao chicken.

In higher stakes cases, such as a medical diagnosis, having the knowledge that AI is being used in that decision might allow you to work with your physician on moving forward. It also may seem useful to have some kind of detailed explanation of the system's internal (causal) structure. Thus, it might seem that high-stakes cases are a kind of equalizer, leading to high utility of both technical and social dimensions. However, in some high-stakes cases, such as healthcare, Technical Transparency of AI is likely of little use in illuminating a chain of reasoning for decisions altogether (London 2019). London warns that requiring AI to be technically transparent is 'unfounded and potentially harmful' because causal knowledge in

medicine is largely incomplete; a search for causal understanding of AI in medicine has limitations and drawbacks.

London's discussion focuses on the medical profession, but a similar argument could be made for any field wherein causal knowledge by the decision-makers may not be sufficiently complete (e.g., psychology, economics, some forms of engineering, politics, etc.). Incomplete causal knowledge by field experts reduces our ability to rely on AI and threatens to create false knowledge (see below). And if we consider fields in which causal knowledge is complete, we must ask ourselves why we are using black-box AI (or AI in general) to begin with, as opposed to using explainable solutions (Rudin and Radin 2019).

Neural networks, the primary black box culprits, are statistical tools designed to find bias in data; it is far more important for a professional to state the biases found in the data than to state the way the bias was found (London 2019). Focusing on the accuracy and utility of the output is often more important than explaining the steps taken to arrive when our causal knowledge is incomplete because, as London said, "When the demand for explanations of how interventions work is elevated above careful, empirical validation, [stakeholders] suffer, resources are wasted, and progress is delayed." Thus, in some fields, particularly those which are high-stakes, it is more effective to focus on Social Transparency to get the job done.

***Regulation*:** There is mounting regulatory pressure for firms to increase transparency, and this compulsion skews towards Technical Transparency.[5] Regulators appear to believe that having visibility and explainability of the causal structures will suffice as meaningful regulation.

[5] The Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) in the United States require high levels of causal transparency in the credit process, and recent executive orders have begun laying the groundwork for transparency initiatives that outline both Technical and Social Transparency (Trump 2020). The European Union has passed regulations claiming that stakeholders have the right to "The logic involved in any automatic personal data processing and...the consequences of such processing" (Regulation 2016). And China has released draft regulations that mention transparency and inspectability (Which at the time of submitting this paper appear to have now been implemented.) (Notice 2021)

But this push for greater technical transparency is hardly justified. At first glance, writing legislation in terms of Technical Transparency does seem to be more rational because it is narrow and rigid. Building regulation in terms of Social Transparency appears to be more complex and less well-defined, requiring complicated social analysis. However, problems arise for Technical Transparency when we attempt to define what exactly we mean to enforce, and its apparent advantages tend to be disadvantageous in practice.

Lawmakers are tasked with the problem of identifying just what kinds of AI should be regulated and how. Regulation built on a framework of Technical Transparency necessarily aims at black-box AI and overlooks less complicated solutions. This is because solutions that are not currently black-boxes need not be regulated (or pressured) to be explainable; they already are. Regulating in terms of Technical Transparency is a narrow approach that has the potential to ignore a large portion of the AI landscape and runs the risk of satiating the hunger of activists before any Social Transparency is realized.

Moreover, when we understand AI as a social entity, it inspires us to regulate AI in terms of its social impact regardless of architectural complexity and opacity. We do not regulate human agents by their neural processes or brain structure, and we regulate them through the abstraction of their actions (Wilson and Shpall 2016). The same would be true of socially regulated AI. Not only does this approach place more bounds on the use-cases of AI, but it is far more easily malleable. There is no guarantee that legislation written for neural networks will apply to technologies ten years ahead, but social rules apply regardless of the underlying agent. This highlights the importance of understanding AI as a broad and socially active technology; it allows us to very easily define the object of our legislation.

In summary, Social Transparency appears to be a more effective and useful treatment to several key ethical points in AI and regulation as a whole. If we have the option, pursuing Social Transparency ahead of Technical Transparency should be the modus operandi. However, there is one more consideration to be made: what if we do not have the option?

Dangers of Technical Transparency

If we cannot have Social Transparency, would having Technical Transparency be better than nothing? Possibly not. Beyond the back and forth of whether to allocate energy into different kinds of transparency, continuing to push for Technical Transparency in itself carries several key risks that could oppose the aims of ethics. Not only does it subtract from spending our energy on more fruitful projects (e.g., Social Transparency), it is potentially harmful in at least four ways, including justified social opaqueness, the creation of false knowledge, justification for prejudice, and impaired system security.

***Justified Social Opaqueness*:** Referring to the previous discussion of regulation, calling for transparency and focusing largely on the technical aspects allows scenarios where firms who claim their models are explainable are able to keep them behind closed doors. The private sector is more interested in Technical Transparency (Jobin et al. 2019), and this is because they believe that if they can say they are following principles, then they will be left alone (Ochigame 2019; Rességuier and Rodrigues 2020). Promotions of Technical Transparency can be a kind of offering by tech firms in hopes of reducing demands for Social Transparency and regulation. Being able to explain why a model malfunctioned may be enough to keep the public complacent.

This kind of complacency is dangerous. Technically transparent systems with no social transparency will plausibly only be investigated after they cause some kind of harm.[6] We could find ourselves in a scenario where AI is only inspected with a warrant once it is implicated. The risk involved in AI is so immense that technically understanding what went wrong may be a day late and a dollar short (Bostrom 2016; Bostrom 2017).

Even if Technical Transparency allows developers to build AI that they are entirely confident will do what they want without malfunction, the discussion around what should be done is still contentiously debated in many decision-making cases[7]—Ethics is an ongoing discussion (Rességuier and Rodrigues 2020). A lack of Social Transparency created by an offering of Technical Transparency risks halting this discussion altogether. It moves the black box problem one layer up from the AI systems to the developers and firms and requires stakeholders to place their full trust in tech companies. But public perception of big tech is on the decline (Auxier 2020; Vogels 2021). If increasing trust is a goal, then we should not be willing to allow an aggressive pursuit of Technical Transparency in this way.

Creation of False Knowledge: Beyond the avoidance of regulation, Technical Transparency could falsely illuminate causal structures which do not exist. London (2019) cites the example of an AI tasked with assessing the risk of death for patients with pneumonia, wherein it was found that patients with asthma have a far lower risk of fatality. A Technically Transparent (i.e., causally transparent) system would show that having asthma is a cause for lower fatality from pneumonia—a result that is counterintuitive and nonsensical. In fact, as London points out, patients with asthma and pneumonia are admitted directly into intensive care,

[6] It may not even be intentionally malicious AI that causes harm; it may be a system with the potential to be malicious without enough eyes on it or a system which we, the public, do not want making a certain kind of decision.
[7] E.g., should an agent driving a car run over a child or hit a tree (if there are no other viable options)?

thus resulting in a higher chance of survival due to the added medical attention. Whereas a socially transparent analysis may have included this information, the raw causal structure leaves us with a misleading and potentially harmful conclusion. If we think that asthma causes a lower chance of mortality in pneumonia patients, and we are using this system to allocate resources with few or no other considerations, we may follow this guidance, not giving those who most need help the attention they require. And while this is just one example, there are plausibly other high-stakes scenarios, especially those where our prior causal knowledge is limited, in which this misguidance may prove indiscernible and dangerous.[8] And while it is difficult to speculate what counterintuitive conclusions an AI might draw, it is conceivable that an AI used in a research setting could have its insights taken at face value. If this were to happen, there is no way to know for certain how far such falsities could permeate before being noticed.

Justification for Prejudice*:* Similar to the creation of false knowledge, another way Technical Transparency could cause harm is by promoting and crystalizing stereotypes. As discussed, there is a common misconception that deep neural networks and other focuses of Technical Transparency can easily illuminate causal structures in the data. However, these systems quite often are built on data that contains inaccurate bias. Finding bias in (presumably) clean data is what machine learning is designed to do (London 2019). But bias can arise from a multitude of factors, not all causal or just. Causal structures can appear where none actually exist. This has the potential to justify long-held prejudices as well as foster novel stereotypes. For example, an AI can claim to find a causal link between a certain facial feature and sexual orientation, yet no such link actually exists (Tomasev et al., 2021).

[8] For example, it is possible AI could assign those in cognitive therapy with suicidal ideation as being at low risk of actually committing suicide. Intuitively, this would be because those with suicidal ideation have greater attention placed on them to ensure they do not harm themselves. But if an AI is administering therapy mainly on its own, it could fail to take the proper steps to ensure the safety of a patient.

While it could be the case that the data had some imperceptible bias that seemed to confer a particular face-sex-orientation combination, this feature is not a causal predictor. Scenarios like this could lead to calcification of harmful stereotypes and could include many other feature relationships.[9] While some false relationships are not directly harmful, those which are can be particularly potent. Although these biases can plausibly be found in statistical analysis, the difference is that some proponents of a particular view could claim that AI reveals that "X is because of Y," not just "X is common amongst Y." The claim is much stronger, so too is its potential for misuse and harm.

Impaired System Security: Technical Transparency could reduce security and robustness. If we pursue the explainability of causal structures, we run the risk of those systems being manipulated more directly. If, for example, one knows that living on Main Street is a cause for approval for a higher loan at a lower interest rate, then it would be most advantageous to pay someone who lives on Main Street to allow you to claim their address (or to even falsely claim you live there). Further, it opens up the possibility for bad actors to make use of systems in a more targeted way, such as if it is uncovered that using specific words (or buying a blue car, or owning a particular brand of laptop) causes one to be deemed worthy of monitoring by the FBI, then that information can be used to evade detection.

It can be argued that Social Transparency is also guilty of reducing security. If the system is technically explainable and is made visible to the entire world, then everyone can view the weaknesses. However, this is overcome by a simple fact of open source development: more eyes on a problem means more eyes to see what is wrong (Bostrom 2017). The most dangerous system is likely to be one that no one knows exists except for a key few. If it is technically

[9] Relationships between: certain addresses and having poor home security, race and creditworthiness, birth dates and medical risks, hobbies and religious affiliation, etc.

transparent, someone on the inside could easily cause harm or otherwise manipulate the system to their advantage. As with many cases, the potential harm of Technical Transparency is only remedied when counterbalanced with Social Transparency.

Conclusion

The pursuit of technical explainability in AI is ethically troubled. In some cases, these troubles appear to be remedied only by an equal or greater pursuit of social openness. And as AI begins to take on more substitutive roles in socially impactful occupations, the need to pursue the most relevant kinds of openness is amplified. Trust, risk mitigation, control, usefulness, and regulation are each better served by pursuing Social Transparency. With further consideration, aspects such as autonomy of those affected by AI, long-term sustainability, fairness, and privacy may plausibly be equally served by social openness and similarly troubled (or unhelped) by technical explainability.[10]

The stakes are high with AI. We may not have the chance to stop rogue systems before they run rampant. Accordingly, we should take extra care to ensure that the principles we enumerate are effective and reduce risk as much as possible. When such care is taken, technical explainability falls short. It seems to do the opposite of the goal, being both less effective and prone to risks. This is a counterintuitive conclusion. Knowing everything about how an AI works should only improve the situation. But in this case, our intuitions appear to fail.

[10] Socially opaque AI could plausibly manipulate stakeholders and affect their autonomy; with the increase of distrust amongst the public and desires for control of AI, long term sustainability might rest on a level of social openness (Fast and Horvitz 2016); while fairness might be informed on one end by technical inspectability, there are also aspects to consider such as outcomes which might be better informed by social analysis (which would necessarily require open access to data and the model); privacy could be benefited similarly to safety, using the principle of more eyes on the problem. These and other considerations could be made in further research.

Bibliography

Asan, Onur, Alparslan Emrah Bayrak, and Avishek Choudhury. 2020. "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians." *Journal of Medical Internet Research* 22 (6): e15154. https://doi.org/10.2196/15154.

Auxier, Brooke. 2020. "How Americans See U.S. Tech Companies as Government Scrutiny Increases." *Pew Research Center* (blog). October 27, 2020. https://www.pewresearch.org/fact-tank/2020/10/27/how-americans-see-u-s-tech-companies-as-government-scrutiny-increases/.

Bostrom, Nick. 2016. *Superintelligence: Paths, Dangers, Strategies*. Oxford, United Kingdom ; New York, NY: Oxford University Press.

———. 2017. "Strategic Implications of Openness in AI Development." *Global Policy* 8 (2): 135–48. https://doi.org/10.1111/1758-5899.12403.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 316–34. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020.

Bringsjord, Selmer, and Naveen Sundar Govindarajulu. 2020. "Artificial Intelligence." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/.

Fast, Ethan, and Eric Horvitz. 2016. "Long-Term Trends in the Public Perception of Artificial Intelligence." *ArXiv:1609.04904 [Cs]*, December. http://arxiv.org/abs/1609.04904.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, Deborah G., and Mario Verdicchio. 2019. "AI, Agency and Responsibility: The VW Fraud Case and Beyond." *AI & SOCIETY* 34 (3): 639–47. https://doi.org/10.1007/s00146-017-0781-9.

Larsson, Stefan, and Fredrik Heintz. 2020. "Transparency in Artificial Intelligence." *Internet Policy Review* 9 (2). https://doi.org/10.14763/2020.2.1469.

London, Alex John. 2019. "Artificial Intelligence and Black-Box Medical Decisions: *Accuracy versus Explainability*." *Hastings Center Report* 49 (1): 15–21. https://doi.org/10.1002/hast.973.

Monett, Dagmar, Colin W. P. Lewis, Kristinn R. Thórisson, Joscha Bach, Gianluca Baldassarre, Giovanni Granato, Istvan S. N. Berkeley, et al. 2020. "Special Issue 'On Defining Artificial Intelligence'—Commentaries and Author's Response." *Journal of Artificial General Intelligence* 11 (2): 1–100. https://doi.org/10.2478/jagi-2020-0003.

Müller, Vincent C. 2021. "Ethics of Artificial Intelligence and Robotics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

"Notice of the State Internet Information Office on the 'Regulations on the Management of Recommendations for Internet Information Service Algorithms (Draft for Solicitation of Comments)' Public Solicitation of Comments." n.d. Accessed November 5, 2021. http://www.cac.gov.cn/2021-08/27/c_1631652502874117.htm.

Ochigame, Rodrigo. 2019. "The Invention of 'Ethical AI': How Big Tech Manipulates Academia to Avoid Regulation." *The Intercept* (blog). December 20, 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/.

"Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1–88 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)." 2016.

Rességuier, Anaïs, and Rowena Rodrigues. 2020. "*AI Ethics Should Not Remain Toothless!* A Call to Bring Back the Teeth of Ethics." *Big Data & Society* 7 (2): 205395172094254. https://doi.org/10.1177/2053951720942541.

Rudin, Cynthia, and Joanna Radin. 2019. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition." *Harvard Data Science Review* 1 (2). https://doi.org/10.1162/99608f92.5a8a3a3d.

Sambasivan, Nithya, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. "Re-Imagining Algorithmic Fairness in India and Beyond." *ArXiv:2101.09995 [Cs]*, January. http://arxiv.org/abs/2101.09995.

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, and et al. 2016. "Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence." https://apo.org.au/node/210721.

Sundar, S Shyam. 2020. "Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII)." *Journal of Computer-Mediated Communication* 25 (1): 74–88. https://doi.org/10.1093/jcmc/zmz026.

Tomasev, Nenad, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. "Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 254–65. Virtual Event USA: ACM. https://doi.org/10.1145/3461702.3462540.

Trump, Donald J. 2020. *Exec. Order No. 13960*. https://www.federalregister.gov/d/2020-27065.

Vogels, Emily a. 2021. "56% of Americans Support More Regulation of Major Technology Companies." *Pew Research Center* (blog). July 20, 2021. https://www.pewresearch.org/fact-tank/2021/07/20/56-of-americans-support-more-regulation-of-major-technology-companies/.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *ArXiv:1711.00399 [Cs]*, March. http://arxiv.org/abs/1711.00399.

Wilson, George, and Samuel Shpall. 2016. "Action." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/action/.

Yampolskiy, Roman V. 2020. "On Controllability of AI." *ArXiv:2008.04071 [Cs]*, July. http://arxiv.org/abs/2008.04071.