Welfarist Moral Grounding for Transparent AI

Devesh Narayanan
National University of Singapore
devesh@u.nus.edu

ABSTRACT

As popular calls for the transparency of AI systems gain prominence, it is important to think systematically about why transparency matters morally. I'll argue that welfarism provides a theoretical basis for doing so. For welfarists, it is morally desirable to make AI systems transparent insofar as pursuing transparency tends to increase overall welfare, and/or maintaining opacity tends to reduce overall welfare. This might seem like a simple – even simplistic – move. However, as I will show, the process of tracing the expected effects of transparency on welfare can bring much-needed clarity to existing debates about when AI systems should and should not be transparent. Welfarism provides us with a basis to evaluate conflicting desiderata, and helps us avoid a problematic tendency to reify trust, accountability, and other such goals as ends in themselves. And, by shifting the focus away from the mere act of making an AI system transparent, towards the harms and benefits that its transparency might bring about, welfarists call attention to often- neglected social, legal, and institutional factors that determine whether relevant stakeholders are able to access and meaningfully act on the information made transparent to produce desirable consequences. In these ways, welfarism helps us understand AI transparency not merely as a demand to look at the innards of some technical system, but rather as a broader moral ideal about how we should relate to powerful technologies that make decisions about us.

CCS CONCEPTS

• **Human-centered computing** → Collaborative and social computing; Collaborative and social computing theory, concepts and paradigms; • **Social and professional topics** → Computing / technology policy; • **Computing methodologies** → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence.

KEYWORDS

Transparency, Welfarism, Moral Theory, AI Ethics

ACM Reference Format:
Devesh Narayanan. 2023. Welfarist Moral Grounding for Transparent AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3593013.3593977

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3593977

1 INTRODUCTION

As opaque AI systems are used to make increasingly morally significant decisions about our lives, there have been mounting calls for AI transparency [29–31, 44]. The exact language varies,[1] but most policy conversations about AI feature, in some form, a demand that one ought to be able to obtain factual, clear, and direct explanations of any decision-making process where an AI system is involved [42]. In recently-published meta-analyses of documents describing governance efforts around 'ethical AI', transparency was found to be among the few principles that were endorsed universally by corporations, policymakers, and academics alike [29, 38, 44]. Few principles for the governance of emerging technologies have been able to claim such unambiguous acceptance.

However, this agreement at the level of principles masks deep confusion about why exactly AI transparency matters morally. Few serious attempts are made to articulate the underlying moral considerations for *why* one ought to care about AI transparency – save the occasional platitude about how 'sunlight is the best disinfectant'. When scholars and policymakers do attempt to justify pursuing transparent AI, their proposals are varied and disparate. Transparency is, for instance, supposed to be instrumental for securing trust in AI, helping decision-makers spot their system's biases and errors, allowing decision-makers to retain meaningful control over AI-supported decisions, and enabling decision-subjects to contest and seek redress for harmful decisions, to name a few. But this only kicks the can down the road. Why do these goals matter morally? What happens when different goals come into conflict with each other? And is transparency sufficient – or sometimes even necessary – for the pursuit of these goals?

Several commentators have picked up on such weaknesses in popular calls for AI transparency, and have raised interesting technical and philosophical challenges against the principle. Transparency is unnecessary, some argue, in cases where AI systems make minor and inconsequential decisions, in cases where making an AI system transparent exposes vulnerabilities for malicious actors to exploit, or in cases where we wouldn't have ordinarily asked for transparency if a human – rather than AI system – was the one making decisions. Some of these challenges are unreasonable, and we should defend the principle against them. Others are reasonable and clarifying, and we should incorporate them to sharpen our understanding of why and how to pursue transparency.

To do all this, we need to locate systematic moral grounding for the principle of AI transparency. I'll argue that such grounding – that consolidates most reasonable intuitions from both sides of the debate – may be found by turning to welfarism. For welfarists, in general, it is morally desirable to make AI systems transparent

[1] Although some scholars have offered taxonomies to separate various terms related to transparency from one another [19, 23, 56, 91], most scholarly and policy writings on the topic use these terms interchangeably. I'll follow this general trend: although I use the terms 'transparency' and 'opacity' throughout, I'll treat them as equivalent to other popular terms such as explainability, (un)intelligibility, and (un)interpretability.

insofar as pursuing transparency tends to increase overall welfare, and/or maintaining opacity tends to reduce overall welfare. I'll show that this process of tracing the expected effects of transparency on welfare can bring much-needed clarity to existing debates about when AI systems should and shouldn't be transparent. Welfarist reasoning provides us with a common basis to evaluate conflicting desiderata, and helps us avoid a common mistake of reifying trust, accountability, and other goals as ends in themselves. And, by shifting the focus away from the mere act of making an AI system transparent, towards the harms and benefits that transparency brings about, welfarists call attention to often-neglected social, legal, and institutional factors that determine whether relevant stakeholders can access and act on information produce desirable consequences.

2 THEORETICAL GROUNDWORK

Applied ethicists turn to moral theory because we want to know facts about which kinds of actions are permissible and impermissible, and perhaps more importantly, *why* these facts obtain. In other words, we want to make sure that our moral judgments are 'well grounded': that they account for the various considerations that compete for our attention when we are thinking clearly about what we ought to do [45]. In cases where there are unclear and/or competing intuitions about how we ought to act, it can be instructive to ground such intuitions in a general moral principle.

My position in this paper is that several persistent confusions in existing debates about AI transparency can be clarified by tracing the effects of providing transparency or retaining opacity on overall welfare – i.e., by 'grounding' the principle in welfarism. This position does not require a full defense of welfarism as a normative theory, nor does it strictly foreclose the possibility of grounding AI transparency by appealing to alternate moral principles. I simply hope to show that welfarist reasoning offers clear and reasonable explanations for the underlying moral considerations behind most popular intuitions about AI transparency, and that my account fares favorably compared to a few other similar attempts to explain why we ought to care about AI transparency. In other words, I do not explicitly argue on independent grounds that welfarism is *true*. I simply wish to show that it is *useful* for clarifying the specific applied ethics debate that this paper is concerned with.

2.1 On 'Welfarism'

Welfarists believe that "[t]he judgment of the relative goodness of alternative states of affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual [welfare] in these states" [78]. This view – following the philosopher Amartya Sen's formulation – entails two broad commitments. First, welfarists believe that welfare is the only property of states of affairs that has intrinsic moral value. Second, welfarists – like some consequentialists – believe that the goodness of a state of affairs scales according to the amount of welfare it contains.[2]

Note that Senian welfarism is not distribution-neutral, and is therefore more restrictive than what consequentialists who care about welfare are committed to accept. Egalitarians or prioritarians who value welfare, for example, might also want to call themselves 'welfare-consequentialists', but Sen's conception implicitly rules this out. I think that this distinction is useful for this project. Since welfare-egalitarians, welfare-prioritarians, and Senian welfarists all endorse different views about how welfare should be distributed, they would appeal to quite different considerations when evaluating the moral desirability of AI transparency, even when all three views converge.[3] This makes it practically unfeasible for my project to treat 'welfarism' as neutral between these views, since I would have to consider various distributional arrangements when evaluating each argument for and against transparency. As such, although I remain open to the possibility of welfare-prioritarian or welfare-egalitarian projects to ground transparency, I will retain Sen's restricted concept of welfarism in this paper.

However, Senian welfarism *is* neutral between different theories of 'welfare'. Welfare is commonly conceptualized in one of three ways. On *hedonist* views, welfare consists of experiencing more pleasure and/or less pain; on *desire-satisfaction* views, welfare consists of getting what you want; and on *objective list* views, welfare consists of achieving certain objectively-specified goods like autonomy, knowledge, love, or virtue [54]. Although hedonists, desire-satisfactionists, and objective list theorists disagree about how to evaluate different states of affairs, within the context of my project, I don't think that their prescriptions for when we should and shouldn't pursue AI transparency would differ very much from one another. Most welfarists have reasonably convergent pre-theoretical intuitions about what constitutes welfare – as Simon Keller (2009) writes, the things that increase one's welfare are those that, intuitively enough, "advance her best interests, or benefit her, or make her life go better, or make things better for her, or make her better off in the most fundamental sense" [49]. It would be more precise, to be sure, for this paper to pick and defend a specific theory of welfare. However, I think that this slight loss in precision is outweighed by the benefits of retaining the appeal of my discussion to a larger group of welfarists. As such, I'll follow Sen in remaining neutral between different theories of welfare.

I hope this account of welfarism is at once general enough to appeal to those who hold a variety of moral views, while specific enough to be useful for effectively intervening in the particular applied ethics debate I'm interested in.[4]

3 THE AI TRANSPARENCY DEBATE, IN WELFARIST TERMS

Transparency has long been considered to be a key enabler of democratic accountability and legitimacy, and a key guardrail against

[2] In this sense, Senian welfarism is a kind of *scalar* consequentialist view, rather than a *maximising* view. Scalar consequentialists believe that rightness and wrongness are a matter of degree, rather than believing that the *only* right thing to do in any situation is that which maximises welfare [81].

[3] Given the massive inequalities of power and informational access that characterize most current AI deployments – and the resultant fact AI systems are most harmful (and most opaque) to those who already tend to be worse off – it is likely that prioritarians, egalitarians and Senian welfarists would converge in many cases. As such, although I'll maintain the Senian conception here, it is likely that welfare-prioritarians and welfare-egalitarians would find much of the discussion in this paper germane to their own attempts to ground AI transparency.

[4] Many moral views accept that welfare matters in some way, even if they disagree that welfarism is the only thing that matters. Some who hold such pluralistic views even call themselves welfarists [62, 82]. As such, it is likely that those who hold non-welfarist moral views (or more narrowly, welfarist views that differ from Senian welfarism) might still find many of the arguments in this paper generally compatible with their views.

arbitrary and unjust exercises of power. For public and private institutions, transparency has come to be seen as a key feature of governance and legal mechanisms to ensure accountability, prevent corruption, improve performance, and increase trustworthiness [67]. Contemporary calls for AI transparency often draw implicit support from this popular acceptance of transparency. AI systems are often seen as comparable to other public or private decision-makers that we already expect transparency from, and hence, calls for their transparency are justified on the same grounds as the transparency regulations we already have.

Although there are very few systematic treatments of the moral considerations underlying AI transparency, those who argue for transparency typically *do* appeal to morally-relevant goals (e.g. trust, accountability, scrutiny) that transparency helps us secure, and those who argue against transparency typically appeal to morally-relevant goals (e.g. accuracy, speed, privacy) that transparency forecloses. The task of this section, therefore, is to show that these various existing moral considerations about transparency are captured – and in many cases, made clearer – by grounding the principle in welfarism. Moreover, it is important to note that welfarist voices have thus far been relatively absent in conversations about AI transparency. By recasting concerns about transparency in welfarist terms, I hope to make it easier for those who are sympathetic to welfarism to join in these conversations.

Per welfarism, it is desirable to make AI systems transparent insofar as pursuing transparency tends to increase overall welfare, and/or maintaining opacity tends to reduce overall welfare. Welfarists therefore treat transparency as an *instrumental* good: i.e., something to be desired not for its intrinsic value, but only for its effects on overall welfare. As such, a welfarist analysis of the transparency debate would begin by tracing how exactly transparency secures or forecloses other moral goods (like trust, accuracy, accountability, etc.), and how exactly these secondary goods are related to overall welfare. I'll show that such analysis is well-suited to help us think clearly about why transparency matters morally.

3.1 Scrutiny

Decisions made by AI systems reflect and reinforce social bias and inequity [3, 15, 20]. In turn, calls for transparency often appeal to the need to identify and eliminate such biases and errors from opaque systems. As the argument goes: if developers are able to access the model's decision-making logic and technical parameters, they would be in a better position to perform routine sanity-checks of the model's logic, identify biases, spurious correlations, distributional drifts, and other such errors, and where possible, design appropriate technical solutions to address these errors [90]. In these ways, transparency could help developers build unbiased, non-discriminatory AI models that work better for everyone.

All else equal, an AI model that is systematically biased against marginalized groups is worse for overall welfare than one that delivers fair outcomes for everyone. In part, this is because the marginal increase in overall welfare from providing favorable outcomes to underserved minorities far exceeds the marginal utility of making those who are already well-served better off. Moreover, as we hear more about the harms of AI systems, public trust in the technology has been dwindling, and people seem less willing to accept and use AI systems even when they could make them better off [33, 51]. If engineers were able to scrutinize and fix these biases as they emerge, and only deploy AI systems that demonstrably work well for those who are most likely to be suspicious of the technology, this could help ensure that people do not end up foreclosing the use of socially-beneficial AI systems.

Crucially, however, these positive effects on welfare are contingent on engineers being able to meaningfully identify and fix an AI model's biases by scrutinizing its inner logic and parameters. It is unclear if this is actually the case. Kaur et al., for instance, studied how AI engineers interact with their model's explanations, and found that engineers tended to take the *mere fact* that a model was transparent as reason to believe that it was unbiased and fair, instead of actually scrutinizing the provided information to come to their own judgments [48]. Providing transparency, here, seemed to make it *less* likely for engineers to scrutinize their models and fix biases. Annany and Crawford [2] offer a sharp diagnosis in this regard: calls for transparency often "[sidestep] the material and ideological complexities and effects of seeing, and [suggest] a kind of easy certainty that knowing comes from looking" [5]. Indeed, it takes certain kinds of special (moral, sociological, political) knowledge to identify biases and discriminatory patterns that might be contained within an AI model's technical parameters, which engineers might not always have [42].

Moreover, engineers need incentives and support for carrying out the tedious, and not always rewarded, work of scrutinizing and debiasing AI models. Given the present political economy under which AI systems are deployed – where organizations are incentivized to launch AI models as quickly as possible, rather than getting them to work safely and robustly – such incentives likely do not exist [6]. Renowned AI engineers have even been fired from their jobs for calling out their companies' harmful AI systems [40, 92]. Welfarists would find it important to think about how incentive structures surrounding the development of AI need to be retooled, so as to meaningfully empower AI engineers to fix harmful AI systems.

Securing these desirable effects of transparency on welfare also requires that those who build and interact with AI systems are able to critically think about the system's logic, to identify biases and other potential sources of error. This may entail building more robust moral education programs for AI engineers, and/or hiring philosophers, STS scholars, and others with relevant expertise to work together with AI engineers in interdisciplinary teams. Moreover, stricter regulations for technology companies, stronger labor protections for technology workers, and perhaps more radically, public ownership and control over AI-developing companies may help establish much-needed structural incentives and protections for AI engineers to scrutinize their models more carefully. Finally, we may also consider shifting the burdens of scrutiny away from engineers, to independent auditors and standard-setting bodies. Such independent auditing practices have worked well for enforcing strict ethical guidelines in other sectors like traffic safety and aviation, and might similarly work as well for AI companies [9, 52]. Welfarists would endorse such policies that might help reliably secure the welfare-benefits of scrutinizing transparent AI systems.

3.2 Contestability

When public and private institutions make decisions on our behalf, it is common to demand redress when these decisions are unfair or otherwise harmful. The ability to contest unfair decisions, as such, is commonly deemed to be "at the heart of legal rights that afford individuals access to personal data and insight into the decision-making processes used to classify them" [63]. Popular calls for 'contestable AI' draw on similar intuitions. Given the biases and inaccuracies that pervade current AI systems, it is widely argued that decision-subjects should have a way to contest unfair AI decisions and seek redress for any harm caused.

Such discussions of contestability feature prominently in discussions about AI transparency. The so-called "right to explanation" of the General Data Protection Regulation (GDPR) is justified on the grounds that transparency is necessary for decision-subjects to learn (a) *that* an AI system was involved in making decisions about them, and (b) *how* these decisions were made, so that they might exercise their right to contest unfair decisions [47]. As the argument goes, if decision-subjects are able to understand the decision-making logic and process of an AI system that makes unfair decisions about them, they would be better able to launch a meaningful challenge against those who developed and deployed the system.

Welfarism would not, in general, frame a 'right to explanation' as if contestability were some primary moral good that we ought to pursue for its own sake. In turn, welfarists would find it harder – relative to, say, deontologists – to defend a right-holder's entitlements to contest and demand compensation for an unfair AI decision. Given that legal efforts to enshrine a right to explanation have been recently making some headway, this is admittedly an important weakness of welfarist approaches to this issue. Nonetheless, welfarists would still be largely supportive of calls for contestability – even if for different reasons. This support draws partly from the fact that compensating those who are harmed by AI systems – especially since they tend to disproportionately be from marginalized and underserved communities – tends to be good for overall welfare. Welfarists would also defend contestability based on the welfare-benefits of increasing democratic oversight over the design and development of AI systems. Most AI contemporary systems are developed by a small number of powerful companies, and in turn, most popular conversations on what needs to be done about AI harms have been captured by these corporate interests [32]. If, however, decision-subjects had the power to contest harmful decisions, and if successful contestations impose significant monetary and/or reputational costs to organizations, this could create incentives for AI developers to put in the work to fix their harmful AI systems, which would be good for overall welfare.

Meaningful contestability, however, requires more than just transparency. It is important to make sure that decision-subjects are readily able to access clear and understandable explanations for what an AI-driven decision-making process looks like. Companies often publish lengthy and jargon-filled technical reports that document the inner workings of their algorithms, or place various bureaucratic, monetary, or technical barriers that obstruct people from accessing relevant information [69]. In such cases, AI systems might be technically 'transparent', but not in a way that makes it feasible for decision-subjects to access and use relevant information to contest unfair decisions. Similarly, relevant legal and institutional procedures to challenge AI systems must also be efficient, accessible, and fair. If corporations, with the support of large and well-resourced legal teams, were able to quash all contestations made against them, this would foreclose any potential welfare-benefits from contestability. For the welfarist, therefore, it is not sufficient for decisions-subjects to merely have the *ability*, in some mere technical sense, to contest decisions – rather, meaningful contestability also requires that contestation must accessible, timely, inexpensive, and fair. Various policy proposals have been offered to this end: including proposals for making AI contestability part of existing consumer protection laws, and for providing free legal counsel and representation to affected decision-subjects [58, 63, 87]. Welfarists would endorse these and other such policies that, in addition to AI transparency, are necessary for realizing the benefits of contestability on overall welfare.

3.3 Accountability

It is often argued that that transparency is necessary for humans to have 'meaningful control' over, and hence be held accountable for, the outputs of their AI systems. Coecklebergh [16] offers perhaps the clearest articulation of this argument, by detailing an 'epistemic condition for accountability'. On his view, it is only when a decision-maker has sufficiently detailed knowledge about how their AI system works, and still chooses to accept its harmful decision anyway, that they can be held accountable for their oversight.

Robust accountability mechanisms in general – and the threat of legal penalties in specific – can serve as powerful incentives for decision-makers (and others in similar positions) to intervene and mitigate an AI system's potential harms before they occur. In such cases, accountability mechanisms could have a net positive effect on welfare, and as such, would be defended by welfarists. However, to secure these welfare-benefits, accountability mechanisms must be *targeted* at those who are best placed to stop an AI system from making harmful decisions. It is often assumed that decision-makers – i.e., the people charged with scrutinizing AI systems to accept or overrule their decisions – are the only ones who occupy these positions. Indeed, one of the main reasons why AI governance policies mandate that every AI system must have a decision-maker (or 'human-in-the-loop') is precisely so that there is always someone to hold accountable [4]. However, targeting accountability mechanisms solely at the decision-maker might not always be the most optimal. Depending on how the organization deploying the AI system is structured, a decision-maker may report to several others hierarchically above them, and these higher-ups would have more power to make policy decisions about using and managing the system's outputs. The decision-maker may also be deciding to accept or overrule their AI system's decisions by following external guidelines and standards of procedure – in which case, the regulatory institutions and standard-setting bodies that develop such guidelines would have significant influence over what decisions are allowed to be executed. Several others besides the decision-maker might be directly or indirectly able to intervene and prevent an AI system's potential harms, and many of these people might have significantly more power and influence to effectively do so. Welfarists would consider these other groups when thinking

about who to make AI systems transparent to, and in turn, who to hold accountable for AI harms. In so doing, we might broaden discussions on AI accountability beyond its current fixation on the 'human-in-the-loop'.

Welfarists would also be concerned about what *kinds* of accountability mechanisms might be best suited to provide the right incentives to those who oversee AI systems. Minimally, accountability requires that the one being held accountable should not be in a position of relative power over the one who is holding them to account, and should not get to dictate the terms under which they would be penalized [66]. These conditions are often unmet. Most of the powerful organizations that currently develop and deploy AI systems have, despite facing increasing backlash for their harmful technologies from regulators and the public, managed to escape relatively unscathed thus far [41, 80]. These organizations are also playing an increasingly dominant role in shaping policy conversations on 'AI ethics': on what kinds of harms warrant regulatory interventions, and what kinds of penalties and incentives are appropriate to facilitate AI accountability [35, 43, 64, 76]. Absent sufficiently powerful institutions to hold technology companies to account, merely making AI systems transparent is unlikely to be sufficient for meaningful accountability [22]. Yet, overly punitive accountability mechanisms might deter smaller technology companies from developing and deploying AI – for fear of being brought to ruin if their AI systems cause harm – which might foreclose potentially beneficial AI systems from being developed in the first place.

In these ways, welfarism calls on us to recenter considerations about the political economy of AI – on how economic and political power in relation to the development and deployment of AI is distributed, and what it would take to incentivize technology companies to intervene in AI harms – when discussing the connections between transparency and accountability. This seems like a deeper and more expansive way of thinking about why transparency and accountability matter morally.

3.4 Trust

Transparency is said to provide an epistemic warrant for trust. For decision-makers, knowledge of the inner workings of their AI system, as empirical studies show, can help them come to well-reasoned judgments about when and why to trust its recommendations [11, 27, 57, 71], and in turn, learn to be more discerning about when such trust is warranted.[5] For decision-subjects, transparency can similarly provide reasons to believe that AI systems are working fairly, robustly, and in their interest [18, 71, 93].

For the welfarist, an AI system is worthy of trust insofar as trusting it tends to increase overall welfare. Welfarists therefore encourage trust in, and the adoption of, AI systems that are designed to achieve ends that are beneficial to overall welfare (say, those used to help cities anticipate and prepare for natural disasters, or even those used simply to help us get better at fun games like chess). And, welfarists encourage mistrust and resistance towards AI systems that are deployed towards harmful ends (say, those used to produce misinformation and 'deepfake' images, or those used to target surveillance and policing towards minorities). As such, for welfarists, the desirability of an AI system's transparency depends on the ends served by this system.

In workplaces, for example, workers might sometimes unduly reject AI systems that make their work easier and more efficient (like scheduling assistants or task management software), often out of misplaced fears that adopting these systems will somehow lead to job loss or pay cuts [21, 59, 83]. In such cases, transparency – both in terms of how the system works, as well as its benefits to workers – may help workers more readily adopt beneficial AI systems. At the same time, there are growing worries that transparency can lead to trust in fundamentally untrustworthy AI systems. Empirical studies have suggested that transparency can lead decision-makers to 'overtrust' their AI systems, often because they take the *mere fact* that a system is transparent (rather than the information made transparent) as reason to trust it [11, 34, 48]. Similarly, there are worries that organizations use 'transparency' as little more than a marketing buzzword – providing "empty explanations as a psychological tool to soothe users" [91]. Surely if an AI system deployed towards actively harmful ends – say, to surveil and police a minoritized community – was made transparent, this by itself shouldn't make it any more worthy of trust and adoption.

Welfarism, therefore, asks us to not treat 'trust' and 'adoption' as unalloyed goods, and rather to only trust those AI systems that effectively help us achieve ends that are beneficial to overall welfare. The welfarist's support for transparency insofar as it leads to trust is similarly conditional. In a world where many AI systems are biased and erroneous, and/or deployed towards harmful ends, mistrust and resistance are often warranted. Yet, extant policy and industry routinely treat 'trust' and 'adoption' as key principles for 'ethical AI', and mistrust as an inconvenience that always needs to be overcome as quickly as possible [14, 51]. Highlighting the need to be more discerning about such uncritical framings of trust is, I suggest, a particularly attractive feature of the welfarist view.

3.5 Double Standards

It is often said that AI decision-making is held to an unrealistically high standard. If a human was asked to provide an explanation for how they came to a certain decision, they would, at best, be able to "identify a few factors relevant to their decision, and offer these factors with a few lines in defense of their putative salience" [95]. One certainly would not be able to provide any detailed information about all the factors relevant to their decision, and how exactly these factors were weighed against each other. However, this kind of information is precisely what we seem to expect when we ask for transparent AI. What might justify such a double standard?

It is true that the kinds of detailed information that we might obtain about a transparent AI system's decision-making logic cannot be reasonably obtained from a human decision-maker. However, even if we *could* somehow make humans fully 'transparent', we often would not care to do so. When we ask a friend to choose a restaurant for dinner, or a taxi driver to pick whichever route to our destination he thinks is best, we do not usually care about the exact inner decision-making logic that led them to make particular recommendations, even when these recommendations turn out to be suboptimal. The same might be said about AI. We routinely use

[5] This seemingly straightforward connection between transparency and trust has been problematized in recent scholarship – see e.g. [65].

AI systems to filter spam in our email inboxes, to generate auto-captions for YouTube videos, and to perform various other such minor tasks. Although such AI systems could be made transparent, it does not seem that we would be particularly interested in scrutinizing their decision-making logics.

Conversely, when people (especially groups of people – like governments or corporations) make important decisions on our behalf, we *do* in fact frequently demand transparency from them. When banks deny our loan applications, when hospitals deny us medical care, or when corrupt government officials pass laws in favor of those who bribe them, for instance, we rightfully expect that our public and private institutions should share information about their internal decision-making processes with us. Indeed, calls for transparency from our public and private institutions have long predated the use of AI in decision-making, and many countries have well-established 'freedom of information' and/or corporate transparency laws for precisely this reason. Calls for AI transparency, similarly, seem most urgent when we discuss the use of AI systems in sensitive, high-stakes decisions (e.g., evaluating bail applications, predicting crime, screening resumes, etc.) compared to relatively inconsequential use-cases. The relevant difference here seems to be this: the more consequential the decision in question – i.e., the more likely it is to create large amounts of welfare or harm – the more urgent are our demands for transparency.

For welfarists, this difference is of key moral significance. When a decision – whether made by a human or an AI system – leads to large amounts of harm, transparency can help us understand the factors that led to this decision, and in turn, how we might avoid such harms in the future. And, when a decision creates large amounts of welfare, transparency about its decision-making process might teach us how to make more such good decisions in the future. On these grounds, powerful decision-makers that tend to make more consequential decisions deserve much more careful scrutiny than their less consequential counterparts.

A small-scale deployment of an AI system bears more resemblance to the case where a minor decision is made on our behalf by some layperson. But large-scale AI deployments that make decisions on behalf of numerous people better resemble the kind of impact on overall welfare that, say, a government might have. In these latter cases, there are clear welfare-benefits of having rigorous checks and safeguards in place to ensure that these decision-makers consistently generate good outcomes, and welfarists would find calls for transparency justified on these grounds.

3.6 Trade-offs

We often incur substantial costs – both technical and monetary – when we try to make AI systems transparent. One such widely-discussed trade-off is between transparency and accuracy. AI engineers have found that when transparency is made an explicit design constraint in AI development, the resulting models tend to perform less accurately than their black-box counterparts [36]. Constraining models to be 'glass-boxes', such that their inner decision-making logic is rendered easily accessible, usually entails choosing simpler linear or decision-tree models that usually fail to meet the performance standards set by more sophisticated – albeit opaque – 'deep learning' models. Ostensibly, the "poor performance [of transparent AI models], and their ability to be well-interpreted and easily explained come down to the same reason: their frugal design" [55].

Affordability, speed, and privacy are other desiderata that AI transparency seemingly needs to be traded against. Significant monetary costs might be incurred when pursuing AI transparency: both for the technical work of developing and applying interpretability techniques to make a model transparent, and for setting up organizational processes to ensure that the model's information is scrutinized and acted upon efficiently. Further, many current deployments of AI are too rapidly-paced for human decision-makers to efficiently scrutinize each decision. The AI systems used to trade stocks and currencies in real-time, or to curate content when we use search engines like Google, for example, are designed specifically to respond as quickly as possible to user inputs. If such AI systems must have their outputs reviewed by human decision-makers, it would become impracticable to deploy AI in time-sensitive use-cases. Finally, Shokri et al. [79] show that statistical information about an AI system's decision-boundaries can be reconstructed to make inferences about the data constituting its training set – in which case, transparency would cut against the demands of privacy.

Welfarists approach trade-offs on a case-by-case basis – comparing the costs and benefits of pursuing transparency, accuracy, speed, and affordability for a particular AI deployment, and picking the option that maximizes overall welfare. This might not always be straightforwardly achieved, since the epistemic obstacles to carrying out such cost-benefit analyses are non-trivial. However, in several cases where the benefits of transparency are either particularly large or negligible, welfarist analysis can be generative. For instance, in cases where there are no human decision-subjects that might be directly affected by an AI system's decisions – such as for AI systems used to filter spam emails or play chess – the welfare-benefits of transparency might be outweighed by the benefits of increased accuracy or speed. Or, in cases where contestability and accountability mechanisms are of critical importance – such as for AI systems used to evaluate parole applications or issue medical diagnoses – the welfare benefits of transparency might outweigh those of competing desiderata. Moreover, even when a complete cost-benefit analysis is difficult, it can still be productive to analyze trade-offs in terms of their welfare-effects. When considerations for and against transparency are recast in terms of welfare, they can be meaningfully compared and aggregated. As such, within the context of specific and clearly-defined AI deployments, welfarists can make reasonable headway by estimating the *relative* welfare-effects of transparency versus other desiderata in terms of a common basis. At a time when some take the mere fact that transparency needs to be traded off against other goods to mean that we should give up on transparency altogether [10, 39, 86], welfarist analysis can offer a more careful approach to thinking through these trade-offs.

Moreover, when welfarists are forced to pick between competing moral goods, we would want to find ways to recover as much of the lost welfare-benefits as possible. This can be feasibly pursued in a few ways. Many trade-offs – especially the much-touted transparency-accuracy trade-off – are relevant only when we pursue a *particular* kind of transparency: i.e., when we try to turn 'black-boxes' into 'glass-boxes' by imposing transparency as a strict

design constraint. This may not always be necessary [56, 75]. Complex AI models can be made *approximately* transparent through various post-hoc interpretability techniques – by testing them on a variety of inputs, or by simulating their decision-making logic using secondary AI models – which can often be enough for developers and users to perform careful scrutiny and debugging. Even weaker kinds of 'transparency' may be achieved simply by letting users know *that* an AI system was involved in making a decision about them, by informing them about who built this system and how it was tested and developed, and/or describing the ethical guardrails and governance mechanisms that were set up for this AI system. These kinds of transparency need not entail a trade-off with speed, accuracy, or affordability, and yet might suffice for ensuring contestability, building trust, and other such desiderata. Finally, the logic of each decision made by a rapidly-paced AI system need not be scrutinized by a human decision-maker in real-time. Decision-logics could be stored in logs, and human reviewers could check these logs at some later time. As long as the relevant welfare-benefits of transparency can still be achieved, welfarists would support such alternate approaches to AI transparency.

3.7 Gaming the System

For transparency to be useful, some argue, the release of information must be disciplined. Some worry that the indiscriminate sharing of information about an AI system's inner logic would enable decision-subjects to accordingly change their behavior to gain undeserved rewards [5]. For instance, transparency in the AI systems used for policing and crime-detection could allow criminals to escape detection entirely, by altering their behavior such that they are not flagged as risky. Similarly, it might seem counterproductive to make an AI system used for hiring workers transparent, since job applicants might adjust their behavior to trick the system into overestimating their suitability for any given job [68]. In such cases, when information made available about AI systems allows for them to be successfully 'gamed', their predictive power is invalidated, and as such, they become largely unusable [56].

There are important cases where such worries about gaming systems reflect serious material harm. Consider an AI system deployed to control traffic in a highly secure military network. If information about this system was made transparent to a malicious group of hackers, they would be able to carry out devastatingly efficient attacks. In this case, provided the system functions well as a black-box, it might be preferable to keep it opaque. If no one, not even those who own and deploy the system, fully understands how the system works, there is no chance for critical security information to fall into the wrong hands. This approach is sometimes called 'security by obscurity', and it can sometimes be the most feasible strategy for reducing the risk of harm [60]. Relatedly, if a social media platform makes its content-ordering algorithms transparent to all, well-resourced malicious actors might learn to better manipulate these platforms into promoting misinformation and harmful content. Well-resourced climate deniers, for instance, have already purchased prominent advertisement space to display misinformation at the top of search results about climate change on Google [84]. If such groups were able to access to the inner workings of Google's content-ordering algorithms, their content promotion efforts could become more clandestine, widespread, and dangerous. In such cases, it would be better for overall welfare to keep AI systems opaque (or at least, to restrict who gets access to this information).

However, in many cases, the harms of gaming AI systems are overstated, and do not outweigh the benefits of transparency. AI-deploying organizations have been known to indiscriminately use the language of 'gaming the system' to chastise user actions that are detrimental to their own material interests (of profit, control, etc.), even if they are otherwise beneficial to overall welfare [17, 70]. A Black woman applying for a loan, say, who knows that her application is likely to be unduly rejected by a biased (but transparent) AI system, might choose to leave her race unspecified, or otherwise adjust her application to trick the algorithm into approving her loan. Or, groups of users on a social media platform might use information about the platforms' content-ranking algorithms to crowd-out and/or downvote hate speech and misinformation. As such cases illustrate, some AI systems produce biased and harmful decisions, and gaming them might help mitigate some of their harms. And, in other cases, AI systems can be gamed in ways that produce benefits for society at large, even if this might sometimes make the organizations that deploy these systems worse off. Welfarists would support transparency insofar as it might enable these kinds of welfare-increasing instances of 'gaming the system'.

The gaming of AI systems can also incentivize developers to find and address critical flaws in their systems. Various counter-strategies have been pursued to obstruct users from gaming their systems, and many of these help make AI systems more robust, fair and efficient [5]. In fact, the growth of an entire sub-field of AI research – i.e., 'adversarial machine learning' – might be partially attributed to attempts to respond to 'adversarial' users exploiting the errors and design flaws of AI systems [12]. By incentivizing the development of robust, efficient, and fair AI, the gaming of AI systems can contribute to increasing overall welfare.

Welfarism, therefore, highlights the need to be discerning about when worries about gaming AI systems count against making these systems transparent. In some cases, AI systems are harmful, and *deserve* to be gamed. In other cases, malicious actors can use information about AI systems to cause serious harms, and some kind of opacity might be necessary to mitigate these harms. And in yet other cases, users can game AI systems in ways that create benefits for everyone. Welfarism asks that we consider the harms and benefits that might be brought about by an AI system, and how these might be attenuated or amplified by users who game the system, when considering whether or not to make it transparent.

4 WHAT WE GAIN FROM WELFARIST GROUNDING

By now, I hope to have shown that welfarism captures most prevailing intuitions about when AI transparency should and shouldn't be pursued. This is the bare minimum that one would expect from any attempt to ground some principle in a moral theory. However, I think that welfarism has more to contribute to the AI transparency debate. To make these contributions clear, it is instructive to compare my view with other attempts to defend AI transparency.

Thus far, to my best knowledge, there have been very few articles primarily devoted to moral grounding for AI transparency. One of these is by Kate Vredenburgh [88], who grounds calls for AI transparency in the requirements of fairness. Specifically, she argues that the fairness of institutions depends partly on the ability of individuals to engage in "informed self-advocacy", and that AI transparency is required for such advocacy. Informed self-advocacy entails the ability of decision-subjects to contest unfair and harmful decisions. More broadly, self-advocacy also entails informed deliberation about whether the rules that AI systems use to make their decisions are reasonable and fair. In both cases, AI transparency is necessary for individuals to fully understand – and in turn, to contest and/or revise – the rules that AI systems judge them by. Drawing on Rawlsian concepts of fairness, Vredenburgh then argues that informed consent is necessary to enable a 'fair basic structure' in democratic societies, and for this reason, we are justified in calling for a *basic right* to AI transparency.

More recently, Seth Lazar [53] defended AI transparency by appealing to its necessity for ensuring that decision-making power is subject to norms of *procedural legitimacy* and *proper authority*. On his view, 'procedural legitimacy' requires that: (a) the power of decision-makers must be limited in well-defined ways over a restricted sphere of activity, (b) decision-makers must use their power according to clear and previously agreed-upon rules, and (c) decision-makers must be held to these rules through mechanisms of contestability and accountability. Decision-makers that have 'proper authority' are those that we – typically through democratic processes – explicitly authorize to make decisions about us on our behalf. As such, in cases where decision-makers use (or are replaced by) AI systems, both procedural legitimacy and proper authority require transparency: so that we can better evaluate decision-makers are acting within the boundaries we set for them, understand whether these boundaries are justified, and evaluate whether we ought to continue authorizing their power.

There have been other important attempts to explicate basic moral goods that transparency enables us to secure. Coecklebergh [16] – who defends transparency for enabling decision-makers to be *answerable* to decision-subjects, Binns [7] – who defends transparency for enabling *public reason* and healthy democratic functioning, and Kim & Routledge [50] – who defend transparency for enabling a right to *informed consent*, are representative examples.

The exact arguments by which these views are developed are complex and subtle, and it would fall outside the scope of this project to discuss each of them at length. Instead, I want to focus more generally on what kinds of questions such views help – and fail to help – us answer about the moral considerations underlying calls for AI transparency, and how welfarism fares in comparison. As such, I'll use these accounts as a foil to make the advantages – as well as weaknesses – of my view easier to see.

4.1 Accounting for Diverse Reasons

Those who argue in favor of AI transparency appeal to a wide range of moral goals that transparency helps us pursue, including scrutiny, contestability, accountability, and trust. Insofar as we agree that these are all reasonable goals, it seems reasonable that any attempt to morally ground AI transparency should have something to say about why each of these goals matters morally. Many existing views, however, isolate one or a few reasons for pursuing transparency, and defend the principle with exclusive reference to these reasons. Vredenburgh's account focuses primarily on contestability and scrutiny; Kim & Routledge's account focuses on informed consent, and Binns and Coecklebergh's accounts generally focus on accountability. Lazar's view is perhaps the most broad-ranging, touching variously on themes of trust, public scrutiny, contestability, and accountability, but even his view misses some important cases.

Consider, for instance, the case where an AI system is made transparent *only* to its developers, who scrutinize its decision-making logic to improve its robustness and overall performance. Intuitively, this seems to be the morally right thing to do – *righter*, anyways, than if the system was made transparent to no one. Welfarists can account for this intuition. But since most other views are chiefly concerned with making an AI system transparent to decision-subjects (or, more broadly, to those who are affected by the system), they might not find such developer use-cases of transparency to be morally relevant. Relatedly, welfarists would find AI transparency desirable not only for the sake of contesting and/or holding accountable *existing* AI systems, but also for guiding the *future* development of AI towards morally desirable ends. Using transparency to identify AI hype and snake oil, and to launch more exacting critiques of the political economy of AI is morally relevant to the welfarist, but not always to those who hold other views.

A related worry is that some of these views – in particular, those advanced by Vredenburgh, Coecklebergh, and Kim & Routledge – focus primarily on the rights held by, and/or obligations owed to, *decision-subjects*. This may not always be the most morally relevant consideration, since the people about whom decisions are made are not always those who are most affected. Consider the case where an AI system is used to pick one among a handful of candidates to become the CEO of a large corporation. Here, the relevant 'decision-subjects' include the selected CEO and the other unsuccessful candidates. But, arguably, the ones who are likely to be most affected are the workers of this corporation and/or the broader community that it serves. It seems important to consider this broader group of stakeholders when evaluating whether we ought to make the AI system transparent: either by focusing on the rights and obligations they are owed (as Lazar's and Binns' views do), or on the potential harms and benefits that they might experience (as my view does).

Moreover, some use-cases of AI do not even have decision-subjects. An AI chess-engine, for instance, makes decisions about chess pieces, not people. Still, making such a system transparent – to show what it was considering when it recommended a move (and rejected other options) – might help us learn new moves and strategies, especially for novices who might not have otherwise seen the advantage of the recommended move. Welfarists would find the prospects of more people learning to play better, more fun chess games to be a morally relevant consideration for making AI transparency, while other views might not.

Any attempt to locate moral grounding for AI transparency should take into account the rich and diverse moral reasons that underlie the principle. Welfarism does so, and this is, in my view, one of its most important advantages.

4.2 Remaining Sensitive to an AI System's Goals

Ensuring the transparency of AI systems that could potentially cause large amounts of harm or benefits (say, those used in critical medical applications, or those used to manage traffic systems in large cities) seems more important than the transparency of small-scale, inconsequential AI systems (say, those used in chess engines, or those used to filter spam emails). Welfarism, as we have seen, reliably captures this intuition. This is because welfarists are interested in how transparency might temper or amplify an AI system's effects on overall welfare: for a harmful AI system, transparency is useful insofar as it can help mitigate these harms, and for a beneficial AI system, transparency is useful insofar as it can help amplify these benefits. And, since welfarism is *scalar*, the greater the effect of transparency in amplifying an AI system's benefits or mitigating its harms, the more it ought to be pursued.

Other views do capture some of these intuitions about why the transparency of some AI systems matters more (or less) than others. On Lazar's view, for instance, the greater the decision-making power and authority we delegate to an AI system, the more its transparency matters. And, on Vredenburgh's view, the greater the potential unfairness that a decision-subject might experience, the more urgent the need for informed self-advocacy, and hence transparency. But there are some cases that these views miss.

Consider, for instance, two otherwise identical AI-based surveillance systems: one used by a school to identify and respond to potential school-shooter threats, and another used by an authoritarian government to monitor a group of rebel citizens. Suppose, further, that if the system was made transparent, those being surveilled would eventually learn to behave in ways that won't be flagged as suspicious by the system. If so, in one case, transparency would enable school shooters to enter schools without detection, and in the other case, transparency would enable rebel groups to act against their authoritarian government more efficaciously. Now, in one way, it seems that it would be better for the authoritarian government's system to be transparent, and for the school's system to remain opaque. But, since there does not seem to be any meaningful difference in the amount of decision-making power delegated to either system, or in the ways in which decision-subjects (here, the people being surveilled) might be treated, it is unclear if Lazar's or Vredenburgh's views can help us separate these cases. The difference, of course, seems to be in terms of what these two systems are being *used for*. Welfarists would have little trouble endorsing transparency in the rebel group case – to potentially secure the welfare benefits of undermining authoritarian regimes, and endorsing opacity in the school shooter case – to potentially escape the terrible harms of a school shooting.

The same AI system might need to be made transparent in one context, and remain opaque in another – and this at least partly turns on the ends that the system is deployed towards. Welfarism can help us retain a sensitivity to these differences.

4.3 Calling Attention to Broader Institutional Factors

It is important to ensure that efforts towards AI transparency are sensitive to the moral goals that transparency is supposed to help us pursue. And, especially in relation to contestability, public scrutiny, and public accountability over AI systems, we want to make sure that transparency is not ineffective against those with the power to withstand visibility. Thus, any attempt to ground the principle should give us reasons for ensuring that transparency is relevant, understandable, and actionable.

Welfarism, as I have argued, does this well. By shifting our focus away from the mere technical act of making an opaque AI system transparent, towards the harms and benefits this transparency might bring about, welfarism is *explicitly* concerned with ensuring that stakeholders can access and meaningfully act on the information they are provided to produce desirable consequences. In the previous section, I have discussed various examples of social, organizational, legal, and technical interventions that welfarists would endorse for the sake of addressing existing barriers to the meaningful use of AI transparency. This list of interventions, however, is far from exhaustive. Welfarism tells us that we should care about relevance, understandability and actionability, but doesn't necessarily tell us how to achieve these goals efficiently. In this regard, we have much to learn from other scholars and policymakers who write on the topic. There is extensive scholarship on how to assess which types of information about AI systems are most relevant to different stakeholders [24, 26, 89, 94], how to present information about a model's decision-making logic in a clear and accessible manner [25, 77, 85], and how to ensure that people have the requisite technical and moral knowledge to make sense of the information made transparent to them [8, 28, 74]. In particular, exceptionally comprehensive accounts of how people might use AI transparency to pursue contestability, accountability, and other such goals, may be found in the writings of Vredenburgh, Lazar, and others who seek to morally ground the principle with reference to one or a few of these goals. Welfarists should engage seriously with these scholars, and where possible, incorporate their perspectives to clarify the various technical, social, legal, and organizational interventions needed for AI transparency to increase overall welfare.

4.4 Welfarism and Longtermism

Thus far, although welfarists and other consequentialists have been contributing to AI ethics debates, our contributions have primarily focused on concerns about the 'existential risk' of 'superintelligent' AI systems. Within certain narrow but quite influential circles, there have been growing concerns about hypothetical futures where AI systems might outperform humans on a wide range of cognitive tasks, and in turn, concerted efforts to ensure that such future superintelligent AI systems will be 'aligned' with human values. However, to many scholars and activists fighting the material harms of contemporary AI systems, such worries about superintelligence are irrelevant at best, and obfuscatory at worst.

Longtermists are often accused of ignoring – or intentionally obfuscating – the harms and biases of actually-existing AI systems, in favor of worrying about imagined harms in undeterminable hypothetical futures [72]. And, since most longtermists are welfarists (or, more broadly, consequentialists), welfarism has come to be tarred by the same brush. I think this is a great pity. There is, as I hope to have shown by now, considerable potential in bringing welfarist perspectives to bear on critical conversations about the contemporary harms of AI systems. Longtermist considerations

might emerge out of welfarism, but scarcely represent the entirety of what the theory has to offer.

It is important to note that longtermists *do*, in fact, write about AI transparency, bias and other such contemporary matters, but adopt quite different framings – and in turn, endorse quite different solutions – compared to others who write about these issues. To illustrate: all welfarists take seriously the use of AI transparency for the sake of guiding the development of future AI systems towards prosocial ends. Longtermists typically frame such concerns in terms of the need to learn how to build develop sufficiently advanced transparency tools that could be used to scrutinise and ensure 'value alignment' in future superintelligent AI systems [13]. Longtermists, therefore, endorse research on AI transparency as a way of improving the technical sophistication of transparency methods as much as possible before we get to 'superintelligence'. But this does not strictly follow from welfarist concerns about the development trajectory of AI. There are several *contemporary* practices surrounding AI development – e.g., the diversion of development efforts towards projects that are more flashy than useful, the spreading of obfuscatory AI snake oil and hype, etc. – that welfarists might reasonably worry about. If welfarists instead called attention to such practices – and in turn, to the need to challenge the political, economic, and social forces that presently direct the development of AI towards unproductive and harmful (even if profitable) ends – our contributions would be much more useful and productive in moving forward contemporary AI ethics debates.

All this is to say: the *framing* of welfare considerations matters. If welfarists hope to be taken seriously in critical scholarly and policymaking conversations on AI ethics, we should shift our focus away from the hypothesized future effects of AI in the long term, towards the current and near-future benefits and harms of actually existing AI systems. To do so, we should seek to *make connections* – to explicate how the recommendations of our moral theory bear on existing worries about AI transparency, rather than trying to make its most implausible conclusions our main selling point (cf. Annex A). We should seek to engage carefully with scholars, policymakers, and activists working on the ethical and social implications of AI, to see how welfarism can help advance their efforts. I hope that my paper might serve as inspiration for fellow welfarists who wish to embark on such projects.

5 CONCLUSION

To summarize: welfarists think that this is desirable to make AI systems transparent insofar as pursuing transparency tends to increase overall welfare, and/or maintaining opacity tends to reduce overall welfare. Much of the argumentative work in this paper has been to sketch the explanatory value of this simple-looking move.

Transparency is said to be desirable for the sake of scrutiny, contestability, accountability, and trust, and for better calibrating our interactions with current and future AI systems. Welfarists take these goals as a starting point in their analyses, and in turn, trace out the ways in which pursuing these goals helps (and sometimes, fails to help) increase overall welfare. This approach, as I have shown, is instructive and clarifying. Sometimes, pursuing these various goals can be morally problematic (e.g., when we *over-trust* AI systems, or when we fixate on the accountability of the 'human-in-the-loop' at the expense of all others), and welfarist reasoning helps us separate such cases from others where the pursuit of these goals leads to welfare-benefits. Welfarism also helps us see that transparency is usually insufficient for securing these welfare-benefits, and in turn, draws attention to frequently neglected social, legal and institutional factors that determine whether relevant stakeholders are able to access and meaningfully act on the information they are provided to produce desirable consequences.

Arguments against transparency are similarly made clearer when recasted in welfarist terms. For welfarists, possible trade-offs with other desirable goods, supposed 'double standards', and instances of 'gaming the system' or misleading transparency do not always result in reductions to overall welfare, and hence, do not always count as reasons against pursuing AI transparency. Transparency skeptics sometimes overplay their hands – counting even the slightest risk of harm as reason to give up on transparency entirely – and welfarism helps guard against this tendency. At the same time, welfarists can account for the fact that, in some contexts – say, in cybersecurity applications, where critical information might fall into the wrong hands, or in real-time warehouse monitoring, where the threat of inaccuracies outweighs the need for transparency – transparency can be more harmful than beneficial. In these ways, by shifting the focus towards the harms and benefits that transparency might bring about, welfarism helps be more discerning about when AI transparency is and isn't desirable.

My account, however, has some critical limitations. Due to space constraints, a discussion of these limitations and how they might be addressed has been moved to Annex A.

On a final note, although demands for AI transparency are important, they are not any *more* important than – or indeed, unrelated to – demands for the transparency of powerful people and institutions who make decisions on our behalf. Transparency is a broader moral, political, and social ideal about how we ought to relate with the people, institutions, and technologies that make decisions about us. It is only when we approach the concept at this level of generality – and in turn, make connections to broader concerns about the ways in which we organize our economies, societies, and political institutions – that we might begin to pin down exactly how and why we ought to pursue AI transparency.

ACKNOWLEDGMENTS

The author of this paper was partially supported by a grant from the NUS Centre for Trusted Internet & Community (Grant No.: CTIC-RP-20-06). The author is also grateful to Neil Sinhababu, Zach Barnett, Abelard Podgorski, Zhi Ming Tan, and David De Cremer for helpful comments on earlier versions of this paper.

REFERENCES

[1] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and the Precise4Q consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 20, 1 (November 2020), 310. DOI:https://doi.org/10.1186/s12911-020-01332-6

[2] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 979–989. DOI:https://doi.org/10.1177/1461444816676645

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Laura Kirchner. 2016. Machine Bias. *ProPublica*. Retrieved February 16, 2022 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Jef Ausloos, Pierre Dewitte, David Geerts, Peggy Valcke, and Bieke Zaman. 2018. Algorithmic transparency and accountability in practice.

[5] Jane Bambauer and Tal Zarsky. 2018. The Algorithm Game. *Notre Dame Law Review* 94, (2018), 49.

[6] Ruha Benjamin. 2020. Race after technology: Abolitionist tools for the new Jim code. (2020).

[7] Reuben Binns. 2018. Algorithmic Accountability and Public Reason. *Philos. Technol.* 31, 4 (December 2018), 543–556. DOI:https://doi.org/10.1007/s13347-017-0263-5

[8] Jason Borenstein and Ayanna Howard. 2021. Emerging challenges in AI and the need for AI ethics education. *AI Ethics* 1, 1 (February 2021), 61–65. DOI:https://doi.org/10.1007/s43681-020-00002-7

[9] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (2021), 2053951720983865.

[10] Andrew Burt. 2019. The AI Transparency Paradox. *Harvard Business Review*. Retrieved April 30, 2023 from https://hbr.org/2019/12/the-ai-transparency-paradox

[11] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, IEEE, Dallas, TX, USA, 160–169. DOI:https://doi.org/10.1109/ICHI.2015.26

[12] Ryan Calo, Ivan Evtimov, Earlence Fernandes, Tadayoshi Kohno, and David O'Hair. 2018. Is Tricking a Robot Hacking? *Tech Policy Lab* (January 2018). Retrieved from https://digitalcommons.law.uw.edu/techlab/5

[13] Stephen Cave and Seán S. ÓhÉigeartaigh. 2019. Bridging near- and long-term concerns about AI. *Nat Mach Intell* 1, 1 (January 2019), 5–6. DOI:https://doi.org/10.1038/s42256-018-0003-2

[14] Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. 2021. Trustworthy AI. In *Reflections on Artificial Intelligence for Humanity*, Bertrand Braunschweig and Malik Ghallab (eds.). Springer International Publishing, Cham, 13–39. DOI:https://doi.org/10.1007/978-3-030-69128-8_2

[15] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[16] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics* 26, 4 (2020), 2051–2068.

[17] Ignacio Cofone and Katherine J. Strandburg. 2019. Strategic Games and Algorithmic Secrecy. DOI:https://doi.org/10.2139/ssrn.3440878

[18] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model User-Adap Inter* 18, 5 (November 2008), 455–496. DOI:https://doi.org/10.1007/s11257-008-9051-3

[19] Kathleen A Creel. 2020. Transparency in Complex Computational Systems. *Philosophy of Science* 87, (January 2020), 568–589.

[20] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved August 2, 2022 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[21] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[22] Cory Doctorow. 2019. Regulating Big Tech makes them stronger, so they need competition instead. *The Economist*. Retrieved August 17, 2022 from https://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead

[23] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[24] Lilian Edwards and Michael Veale. 2017. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.* 16, (2017), 18.

[25] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, Yokohama Japan, 1–19. DOI:https://doi.org/10.1145/3411764.3445188

[26] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 449–466. DOI:https://doi.org/10.1007/978-3-030-60117-1_33

[27] Warren J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol.* 34, 4 (December 2021), 1607–1622. DOI:https://doi.org/10.1007/s13347-021-00477-0

[28] Casey Fiesler, Natalie Garrett, and Nathan Beard. 2020. What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ACM, Portland OR USA, 289–295. DOI:https://doi.org/10.1145/3328778.3366825

[29] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020–1 (2020).

[30] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[31] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach (Dordr)* 28, 4 (2018), 689–707. DOI:https://doi.org/10.1007/s11023-018-9482-5

[32] Timnit Gebru. 2021. For truly ethical AI, its research must be independent from big tech. *The Guardian*. Retrieved May 3, 2022 from https://www.theguardian.com/commentisfree/2021/dec/06/google-silicon-valley-ai-timnit-gebru

[33] Crystal Grant and Kath Xu. 2021. Public Trust in Artificial Intelligence Starts With Institutional Reform | News & Commentary. *American Civil Liberties Union*. Retrieved August 5, 2022 from https://www.aclu.org/news/national-security/public-trust-in-artificial-intelligence-starts-with-institutional-reform

[34] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (November 2019), 1–24. DOI:https://doi.org/10.1145/3359152

[35] Adam Greenfield. 2017. *Radical technologies: The design of everyday life*. Verso Books.

[36] David Gunning. 2017. Explainable Artificial Intelligence (XAI). In *DARPA/I20 Project*.

[37] Axel Haenen. 2020. AI transparency in financial services. *Accenture Insights*. Retrieved September 26, 2022 from https://www.accenture.com/nl-en/blogs/insights/ai-transparency-requirements

[38] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30, 1 (2020), 99–120.

[39] Karen Hao. 2020. The messy, secretive reality behind OpenAI's bid to save the world. *MIT Technology Review*. Retrieved April 30, 2023 from https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/

[40] Karen Hao. 2020. We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*. Retrieved September 21, 2022 from https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[41] Karen Hao. 2021. Inside the fight to reclaim AI from Big Tech's control. *MIT Technology Review*. Retrieved August 17, 2022 from https://www.technologyreview.com/2021/06/14/1026148/ai-big-tech-timnit-gebru-paper-ethics/

[42] Tomasz Hollanek. 2020. AI transparency: a matter of reconciling design with critique. *AI & SOCIETY 2020* (2020), 1–9. DOI:https://doi.org/10.1007/s00146-020-01110-y

[43] Lily Hu. 2021. Tech Ethics: Speaking Ethics to Power, or Power Speaking Ethics? *Journal of Social Computing* 2, 3 (September 2021), 238–248. DOI:https://doi.org/10.23919/JSC.2021.0033

[44] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[45] Julian Jonker. 2020. Generic Moral Grounding. *Ethic Theory Moral Prac* 23, 1 (February 2020), 23–38. DOI:https://doi.org/10.1007/s10677-020-10074-3

[46] Shelly Kagan. 1992. The structure of normative ethics. *Philosophical perspectives* 6, (1992), 223–242.

[47] Margot E Kaminski. 2021. The right to explanation, explained. In Sharon Sandeen, Christopher Rademacher and Ansgar Ohly (eds.). Edward Elgar Publishing, 22.

[48] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 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.

[49] Simon Keller. 2009. Welfarism. *Philosophy Compass* 4, 1 (2009), 82–95. DOI:https://doi.org/10.1111/j.1747-9991.2008.00196.x

[50] Tae Wan Kim and Bryan R. Routledge. 2021. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly* (2021), 1–28. DOI:https://doi.org/10.2139/ssrn.3716519

[51] Bran Knowles and John T. Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Virtual Event Canada, 262–271. DOI:https://doi.org/10.1145/3442188.3445890

[52] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, Janet Adams, Christina Hitrova, Jeremy Barnett, Parashkev Nachev, David Barber, Tomas Chamorro-Premuzic, Konstantin Klemmer, Miro Gregorovic, Shakeel Khan, and Elizabeth Lomas. 2021. *Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms*. Social Science Research Network, Rochester, NY. DOI:https://doi.org/10.2139/ssrn.3778998

[53] Seth Lazar. 2022. Legitimacy, Authority, and the Political Value of Explanations. *arXiv* (2022), 21.

[54] Eden Lin. 2022. Well-being, part 2: Theories of well-being. *Philosophy Compass* 17, 2 (2022), e12813. DOI:https://doi.org/10.1111/phc3.12813
[55] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (December 2020), 18. DOI:https://doi.org/10.3390/e23010018
[56] Zachary Lipton. 2019. The Mythos of Model Interpretability. *ACMQueue* 16, 3 (2019). Retrieved June 1, 2021 from https://queue.acm.org/detail.cfm?id$=$3241340
[57] Joy Lu, Dokyun Lee, Tae Wan Kim, and David Danks. 2019. Good Explanation for Algorithmic Transparency. *Available at SSRN 3503603* (2019).
[58] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 106:1-106:25. DOI:https://doi.org/10.1145/3449180
[59] McKinsey. 2018. Adoption of AI advances, but foundational barriers remain | McKinsey. Retrieved September 23, 2021 from https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain
[60] Rebecca T. Mercuri and Peter G. Neumann. 2003. Security by obscurity. *Commun. ACM* 46, 11 (November 2003), 160. DOI:https://doi.org/10.1145/948383.948413
[61] Sachin Modgil, Rohit Kumar Singh, and Claire Hannibal. 2021. Artificial intelligence for supply chain resilience: learning from Covid-19. *The International Journal of Logistics Management* (2021).
[62] Andrew Moore and Roger Crisp. 1996. Welfarism in Moral Theory. *Australasian Journal of Philosophy* 74, 4 (1996), 598–613. DOI:https://doi.org/10.1080/00048409612347551
[63] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*. Social Science Research Network, Rochester, NY. DOI:https://doi.org/10.2139/ssrn.3311894
[64] Luke Munn. 2022. The uselessness of AI ethics. *AI and Ethics* (2022), 1–9.
[65] Devesh Narayanan and Zhi Ming Tan. 2023. Attitudinal Tensions in the Joint Pursuit of Explainable and Trusted AI. *Minds and Machines* (2023), 1–28.
[66] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2022. Accountability in Artificial Intelligence: What It Is and How It Works. DOI:https://doi.org/10.2139/ssrn.4180366
[67] Onora O'neill. 2006. Transparency and ethics of communication. In *Transparency: The Key to Better Governance?* British Academy.
[68] Arielle Pardes. 2022. How Job Applicants Try to Hack Résumé-Reading Software. *Wired*. Retrieved August 24, 2022 from https://www.wired.com/story/job-applicants-hack-resume-reading-software/
[69] Frank Pasquale. 2015. *The black box society*. Harvard University Press.
[70] Caitlin Petre, Brooke Erin Duffy, and Emily Hund. 2019. "Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility. *Social Media + Society* 5, 4 (October 2019), 2056305119879995. DOI:https://doi.org/10.1177/2056305119879995
[71] Wolter Pieters. 2011. Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology* 13, 1 (2011), 53–64. DOI:https://doi.org/10.1007/s10676-010-9253-3
[72] Kelsey Piper. 2022. There are two factions working to prevent AI dangers. Here's why they're deeply divided. *Vox*. Retrieved September 3, 2022 from https://www.vox.com/future-perfect/2022/8/10/23298108/ai-dangers-ethics-alignment-present-future-risk
[73] Stephan Raaijmakers. 2019. Artificial intelligence for law enforcement: challenges and opportunities. *IEEE security & privacy* 17, 5 (2019), 74–77.
[74] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), Association for Computing Machinery, New York, NY, USA, 515–525. DOI:https://doi.org/10.1145/3442188.3445914
[75] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[76] Henrik Skaug Sætra, Mark Coeckelbergh, and John Danaher. 2022. The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. *AI and Ethics* 2, 1 (2022), 15–27.
[77] J. Schoeffer and N. Kuehl. 2021. Appropriate Fairness Perceptions? On the Effectiveness of Explanations in Enabling People to Assess the Fairness of Automated Decision Systems. 153–157. DOI:https://doi.org/10.1145/3462204.3481742
[78] Amartya Sen. 1979. Utilitarianism and welfarism. *The journal of Philosophy* 76, 9 (1979), 463–489.
[79] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the Privacy Risks of Model Explanations. DOI:https://doi.org/10.48550/arXiv.1907.00164
[80] Tom Simonite. 2020. The Dark Side of Big Tech's Funding for AI Research. *Wired*. Retrieved August 17, 2022 from https://www.wired.com/story/dark-side-big-tech-funding-ai-research/
[81] Neil Sinhababu. 2018. Scalar consequentialism the right way. *Philosophical Studies* 175, 12 (2018), 3131–3144.
[82] L. W. Sumner. 1999. *Welfare, Happiness, and Ethics*. Oxford University Press, Oxford. DOI:https://doi.org/10.1093/acprof:oso/9780198238782.001.0001
[83] Yuliani Suseno, Chiachi Chang, Marek Hudik, and Eddy S. Fang. 2022. Beliefs, anxiety and change readiness for artificial intelligence adoption among human resource managers: the moderating role of high-performance work systems. *The International Journal of Human Resource Management* 33, 6 (March 2022), 1209–1236. DOI:https://doi.org/10.1080/09585192.2021.1931408
[84] Hiroko Tabuchi. 2017. How Climate Change Deniers Rise to the Top in Google Searches. *The New York Times*. Retrieved August 24, 2022 from https://www.nytimes.com/2017/12/29/climate/google-search-climate-change.html
[85] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. 1–13.
[86] James Vincent. 2023. OpenAI co-founder on company's past approach to openly sharing research: "We were wrong." *The Verge*. Retrieved April 30, 2023 from https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview
[87] Kaitlyn Vredenburgh. 2019. Explanation and Social Scientific Modeling. *Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences*. (2019), 134.
[88] Kaitlyn Vredenburgh. 2021. The Right to Explanation. *Journal of Political Philosophy* (2021). DOI:https://doi.org/10.1111/jopp.12262
[89] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.
[90] Daniel S. Weld and Gagan Bansal. 2018. Intelligible Artificial Intelligence. *arXiv: 1803.04263 [cs]* (October 2018). Retrieved March 2, 2022 from http://arxiv.org/abs/1803.04263
[91] Adrian Weller. 2017. Transparency: Motivations and Challenges. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11700 LNCS, (July 2017), 23–40.
[92] Julia Carrie Wong. 2020. More than 1,200 Google workers condemn firing of AI scientist Timnit Gebru. *The Guardian*. Retrieved September 21, 2022 from https://www.theguardian.com/technology/2020/dec/04/timnit-gebru-google-ai-fired-diversity-ethics
[93] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2017. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? (2017), 13.
[94] John Zerilli. 2022. Explaining Machine Learning Decisions. *Philosophy of Science* 89, 1 (January 2022), 1–19. DOI:https://doi.org/10.1017/psa.2021.13
[95] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology* 32, 4 (2019), 661–683.

A LIMITATIONS OF WELFARIST MORAL GROUNDING

One may reasonably expect a project aiming to secure moral grounding for the principle of AI transparency to provide clear recommendations and a fully specified trade-off schedule that might help us determine whether we should (or should not) make specific AI systems transparent. But such expectations – as the reader may have already surmised from the discussion in the main body of this paper – are likely to be frustrated.

In some ways, these limitations reflect a general worry about the applicability of moral theory to debates in applied ethics. As Shelly Kagan writes, we typically expect foundational moral theories to "illuminate how the various factors interact in determining the moral status of an act, [explain] which factors outweigh the others in cases of conflict . . . and provide and vindicate the tradeoff schedule in complex cases involving conflicting factors. Of course in practice foundational theories are virtually never worked out in this kind of detail" [46]. In my view, it is still generative to draw on theory to clarify confusions, explicate why certain moral facts obtain, and generally move applied ethics debates towards *reflective equilibrium*. However, I appreciate that such moves may be unsatisfactory to those seeking explicit recommendations about whether and how to pursue the transparency of specific AI systems.

As we have seen, welfarism can offer clear and systematic procedures for thinking through and comparing moral considerations on both sides of the AI transparency debate. Actually carrying out these procedures to completion, however, is another matter. When evaluating the desirability of making any specific AI system transparent, welfarists need to determine (a) the relevant goals that transparency secures and forecloses, (b) the welfare-effects of each of these goals, and in turn, the aggregate effect of the system's transparency on welfare, and (c) what else, besides transparency, is needed to secure these welfare-benefits. Such questions are difficult – if not wholly impossible – to determine completely. AI is a complicated and nascent technology with few comparable precedents, and as a result, the downstream consequences of making AI systems transparent can often be difficult to anticipate. To make matters worse, the technology is surrounded by hype and overinflated expectations, and if one were to uncritically accept this hype in their welfarist calculus, they would likely be led astray. These epistemic difficulties place serious constraints on the applicability of welfarist theorizing to debates about AI transparency.

However, it is important to note that debates about transparency are fairly well-defined. As calls for AI transparency have gained momentum, there has been an explosion in technical, legal, empirical and policy research on accessing the inner logics of complex ML models, making this transparency accessible to relevant stakeholders, and establishing viable pathways for stakeholders to meaningfully use transparency to further their goals. Several scholars and practitioners have also written about AI transparency considerations within specific domains: including medicine, logistics management, financial services, and law enforcement, to name a few [1, 3, 37, 61, 73]. The key task for the welfarist, therefore, is to recast *known* moral considerations about its transparency in terms of welfare and evaluate their relative trade-offs, rather than trying to come up with welfare estimates from scratch. This task is not trivially accomplished, but it is at least *tractable*.

For instance, when a staunch advocate of model accuracy is pressed to recast their concerns in terms of welfare, they might see that, in some contexts, incremental reductions in overall welfare resulting from lower accuracy might not sufficiently override the welfare benefits of, say, contestability. Or, when a staunch advocate for having a 'human-in-the-loop' to hold accountable for AI decisions is similarly pressed to recast their concerns, they might realize that in some cases, such accountability does not do very much to improve overall welfare. These are simple moves, but they can be quite valuable for helping those on different sides of the debate to interact meaningfully with one another.

As such, even when it is impossible to produce precise estimates of the long-term utilities and disutilities of AI transparency, welfarist moral grounding can still be analytically useful. Welfarist analysis is most valuable for identifying which among our existing concerns and priorities are worth acting on, and which kinds of research and policymaking efforts on the topic are worth pursuing further. This value can be realized when we engage deeply with the concerns, priorities, and research findings of those who are already working on the topic. Put simply: for welfarists to provide useful moral grounding for AI transparency, we must ourselves remain grounded.