AI and the expert; a blueprint for the ethical use of opaque AI

Penultimate draft
please do not quote without permission

Amber Ross
University of Florida
amber.ross@ufl.edu

Abstract

The increasing demand for transparency in AI has recently come under scrutiny. The question is often posted in terms of "epistemic double standards", and whether the standards for transparency in AI ought to be higher than, or equivalent to, our standards for ordinary human reasoners. I agree that the push for increased transparency in AI deserves closer examination, and that comparing these standards to our standards of transparency for *other* opaque systems is an appropriate starting point. I suggest that a more fruitful exploration of this question will involve a *different comparison class*. We routinely treat judgments made by *highly-trained experts in specialized fields* as fair or well-grounded even though—by the nature of expert/layperson division of epistemic labor—an expert will not be able to provide an explanation of the reasoning behind these judgments that makes sense to *most other people*. Regardless, laypeople are thought to be acting reasonably—and ethically—in deferring to the judgment of experts that concern their area of specialization. I suggest that we reframe our question – regarding the appropriate standards of transparency in AI as one that asks when, why, and to what degree it would be ethical to accept *opacity* in AI. I argue that our epistemic relation to certain opaque AI models may be relevantly similar to the layperson's epistemic relation to the expert, such that the successful expert/layperson division of epistemic labor can serve as a blueprint for the ethical use of opaque AI.

Keywords:
AI Ethics, Opacity, Transparency, Explicability, Social Epistemology, Expert Testimony

Introduction

Does the widespread demand for increased transparency in AI impose an epistemic double standard on the judgments made by AI models? And if so, are those double standards justified? Should we hold AI models to the same standards of transparency that we hold an ordinary human reasoner? These questions are beginning to receive attention in the AI ethics literature, but to date there is minimal consensus. Zerilli et al. (2018) argue that much of our current proposed regulations would hold AI models to higher than normal—and higher than necessary—standards of transparency. Günther & Kasirzadeh (2022) hold that, while there may be a double standard for ordinary human judgments and judgments made by AI models, those heightened standards for AI are appropriate.

Though they disagree on what the standards for AI transparency *ought* to be, all parties seem to accept that the standards to which they should be *compared* are our standards for transparency in the judgments of *ordinary human reasoners*. This makes sense, insofar as one's own decision-making process is thought to be transparent to oneself, while the reasoning of other minds is notoriously opaque. And in high-stakes decisions, or contexts in which *fairness* is an issue, we certainly require at least some degree of explanation or transparency before we will accept a person's judgment as fair and well-grounded. Though we may not demand a *full* accounting of the reasoning process that ordinary humans engage in when they make these judgments, our standards require that, at minimum, they ought to be able to provide an explanation of their reasoning that *makes sense* to most other people.

While I agree that the widespread push for increased transparency in AI deserves closer examination and that comparing these to our standards of transparency for *other* opaque systems is an appropriate starting point, I believe that a more fruitful exploration of this question will involve a *different comparison class*. While our most *ubiquitous* standards of transparency are those that apply to ordinary human reasoners making ordinary decisions, there is another familiar class of judgments to which these ordinary standards of transparency do not apply. We routinely treat judgments made by *highly-trained experts in specialized fields* as fair or well-grounded even though—by the nature of expert/layperson division of epistemic labor—an expert will not be able to provide an explanation of the reasoning behind these judgments that makes sense to *most other people*. Despite this fact, most other people (those who are not experts in the particular specialized field) would be acting reasonably—and ethically—in deferring to the judgment of experts regarding matters that concern their area of specialization. I suggest that we might make progress on questions regarding the appropriate standards of transparency in AI by reframing the question as one that asks when, why, and to what degree it would be ethical to accept *opacity* in AI. As I will argue, our relation to some opaque AI models may be sufficiently similar to the ordinary layperson's relation to the specialized expert such that analyzing the successful expert/layperson relation may provide us with a blueprint for how to best utilize opaque AI systems, both practically and ethically.

The general organization of this paper will be as follows: In section 1, I will discuss the general value of allowing for the kind of opacity that exists in the expert/layperson relation. In

section 2, I will address the value of transparency in decision-making, focusing on automated decision makers (ADMs) and the problem of bias in machine learning. In section 3, I will explore areas of ethical concern *beyond* bias. Fairness is one value among many that must be considered when developing guidelines for the ethical use of AI. I believe an overly concentrated focus on the problem of bias in AI has drawn our attention away from other values that need to be considered in a full-cost accounting of our use of AI. It is the presence of these additional considerations that show why, in certain cases, allowing for opacity in AI models may be ethically preferable to a constant pursuit of transparency. In section 4, I will argue that the call for transparency in AI is mainly in service of a separate end—that transparency serves as a proxy for the trustworthiness of opaque processes, and increasing transparency aims at establishing appropriate levels of trust between stakeholders and opaque AI models. If this is correct, we may be ethically permitted to utilize opaque AI models provided that this trust and trustworthiness can be established through alternate means. In section 5, I will give an overview of several fundamental features of the expert/layperson relation and make a case for the possibility that the relation between stakeholders and opaque AI models could display these features as well. These features will provide a skeletal blueprint for the ethical use of opaque AI. In section 6, I will suggest preliminary guidelines for evaluating contexts in which it may be ethical to employ opaque AI models, consistent with the blueprint adapted from the successful expert/layperson relation.

1. The value of harnessing opaque processes

As a society, we reap enormous benefits from relying on—or deferring to—expert judgments, especially in high-stakes contexts. Our division of epistemic labor allows laypeople to benefit from the knowledge and judgments of specialized experts without understanding *how* the experts arrived at these judgments nor *why* those judgments are justified. Discovering how to effectively utilize this division of epistemic labor is the very foundation of scientific progress.

Our reliance on opaque expert reasoning is so common that it usually passes without our notice. It may be as trivial as relying on the weather forecast when planning a vacation, or as significant as deciding whether to evacuate our homes (risking our lives and livelihoods) because we know we are in the path of a hurricane. In modern society, one doesn't need to understand the nature of carbon monoxide or nuclear reactions to know that certain levels of CO in the home can be deadly, or that certain nuclear power plants are safe to live near. We can make ethically responsible decisions, including high-stakes decisions, without fully understanding the reasoning process on which we are basing our decision, because it is both epistemically and ethically responsible for us to defer to experts in these matters.

For the vast majority of society, the evidence and reasoning processes of any expert in a specialized field is *opaque,* a genuine "black box". Though it is often in our best interest to defer to these experts' judgments, in doing so we are accepting the outcome of a process that we are aware we do not understand. We– individuals who are not experts in a particular specialized field– can *know* far more than we have the capacity to *understand,* because relying on expert opinion is a reliable way to build knowledge and an ethically responsible way to decide how to act. A medical

expert can only make their reasoning and evidence understandable to a layperson *to a certain degree*; for that reasoning to be transparent to the patient, the patient would need to undergo training similar to that which the doctor underwent to become an expert in their field. This is obviously impractical and undesirable. Instead, we routinely rely on reasoning that we do not understand—especially in high-stakes situations—and this practice is indispensable to modern life. We defer to the judgments of medical doctors, structural engineers, epidemiologists, meteorologists, and computer scientists on a daily basis, and we do so precisely because *we know we do not know* what qualifies as good evidence or good reasoning in these highly specialized fields.

Just as human expertise is most useful in areas where sound judgments require extended and complex training in specialized fields (making the required reasoning opaque to most), AI is most useful in areas where its speed and capacity for data processing greatly surpasses human abilities—the same factors that make certain AI models opaque. And just as the judgment of experts is most valuable in high-stakes situations, the *maximal benefit* we can derive from AI will be in its application to areas that are *central to human welfare* (areas such as health, agriculture, climate, and public safety). The power of AI is a double-edged sword. Its extraordinary speed and unconventional data processing methods are the same factors that can make the most powerful AI opaque to its users and stakeholders, creating ethical concerns regarding whether it *ought* to be used in the very areas in which it could potentially provide the most benefit. The more knowledge we are ethically required to have regarding how an AI model works when it operates in a particular domain, the less likely it is that we will be ethically permitted to use AI applications in that domain.

2. Opacity and the problem of algorithmic bias

The call for transparency in AI aims at safeguarding and improving human welfare—in particular, by protecting vulnerable groups who are most often harmed by opaque AI applications and marginalized in AI development. This goal is and should be a top priority in AI regulation. The speed and processing power of AI not only comes at an epistemic cost; as we have learned, our limited epistemic access to certain AI models can bring with it ethical costs as well. In 2016 investigative journalists at ProPublica published an article that exposed apparent racial bias in the popular risk-assessment software COMPAS, used to aid judicial decision-making regarding individuals' risk of recidivism and eligibility for parole. In 2018, Reuters[1] revealed that the AI hiring algorithm in development at Google showed a strong gender bias.

The push to integrate these ADMs into areas such as recidivism risk assessment, loan approval, and hiring practices, has exposed a tension between two worthwhile goals: (i) increased efficiency in important decision-making processes and (ii) protecting individuals' rights by ensuring such decisions are based only upon ethically appropriate considerations. This tension can become more problematic when the AI models involved are opaque—when the methods by which

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

the AI arrives at a decision cannot be tracked by the relevant parties, whether AI practitioner or stakeholder.

The most powerful AI models—such as deep learning models and models involving vast parameters—are also the least comprehensible. While the engineers involved in creating ADMs like COMPAS may be aware of the content of the training dataset and the parameters at the time of use, the precise role these play in generating the ADM's output often remains unknown. For very complex models, there may be no human (neither AI practitioner nor stakeholder) who understands the actual relevance of each datum to the ADM's eventual prediction. As Riberio (2016) writes, "…if hundreds or thousands of features significantly contribute to a prediction, it is not reasonable to expect any user to comprehend why the prediction was made, even if individual weights can be inspected" (Section 2). Characteristics on which we generally believe it would be unethical to base such decisions—such as an individual's race or sex—may play a role in generating the ADM's decision without our knowledge. Even when such protected information is explicitly eliminated from the dataset, opaque AI models may still display incomprehensible discrimination or 'prejudice by proxy.'[2] An ADM may discover a highly efficient method that utilizes a combination of factors (such as zip code and *alma mater*) in such a way that the output is tantamount to a judgment based on race. The more opaque an AI model, the less certain we can be that the model will be adequately unbiased in its assessment.

In response to the problems that can be generated by opaque AI models, there has been a general push for increasing transparency in AI. Governing bodies, technology watchdog groups, and ethicists have made transparency a priority in AI regulations. The European Commission's 2019 Ethics Guidelines for Trustworthy AI identifies transparency as its fourth out of seven key requirements that AI systems should meet. In January 2020, the White House released its first guidelines for AI regulation which, although they are limited to the private sector and do not mention *transparency* verbatim, do include "trustworthiness," which is intimately connected to the value of transparency. Similarly, The Future of Life Institute explicitly includes two transparency-related items in their (2017) account of the general Principles of AI.[3]

Corporations such as Google and Microsoft have publicly acknowledged the importance of transparency in AI as well. As Microsoft CEO Satya Nadella stated in 2016, "We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence… People should have an understanding of how the technology sees and analyzes the world." And in the framework for a 'Good AI Society', Floridi et al. (2018) call for enhanced explicability in AI when AI is involved in *socially significant decisions*. "Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences" (p.702). The consensus that seems to have emerged in response to the opacity problem has been to treat transparency in AI as valuable

[2] See Barocas (2018)

[3] These principles concern *failure transparency* (if an AI system causes harm, it should be possible to ascertain why), and *judicial transparency* (any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority).

in and of itself, and that the overall benefit we gain from AI increases as transparency increases. That is, we are better off *ethically* the more transparent we make our AI models.

3. Ethically Significant Contexts, - concerns *beyond* bias and fairness

Not all uses of opaque AI give rise to ethical concerns. There are many contexts in which the opacity of an AI model is insignificant simply because we consider the consequences of decisions made in those areas to be trivial. Intuitively, if certain activities genuinely qualify as "for entertainment purposes only," such a context would be trivial, or at least not *ethically* significant. In the most general terms, for a context to be ethically significant the consequences of actions or decisions in that context must at minimum carry a risk of harm (where harm is very broadly construed).[4]

Robbins (2019) is skeptical of the call for transparency in AI, and suggests that while the use of opaque AI is ethically permissible in trivial contexts and certain non-trivial contexts (which he groups together as 'neutral contexts'), it should *not* be allowed to operate in what he labels 'morally sensitive contexts'.

Robbins intends this division between morally sensitive contexts and 'neutral contexts' to largely map onto the distinction between contexts in which we intuitively feel comfortable with the use of opaque AI and contexts in which this opacity seems potentially problematic. Commonly identified ethically problematic contexts of use are those such as judicial sentencing (Berk et al. 2016; Barry-Jester et al. 2015), predictive policing (Ahmed 2018; Ensign et al. 2017; Joh 2017; O'Neil 2016) and medical diagnosis (de Bruijne 2016; Dhar and Ranganathan 2015; Erickson et al. 2017). He writes,

> One reason that using inexplicable decisions in morally sensitive contexts like the ones listed above is wrong is that we must ensure that the decisions are not based on inappropriate considerations… Combine this fact with using ML algorithms for decisions that have moral significance (i.e. decisions which could result in harm-broadly construed to include rights violations) and we have an ethically problematic situation. An algorithm used, for example, to accept or reject your loan request will significantly affect you. A rejection could cause you and your partner significant distress and change the course of your life. (Robbins, 2019, p. 498)

Robbins's analysis seems to suggest that there are two features of a context which together make it a *morally sensitive* context. One concerns fairness. The other is magnitude of impact, or whether it is a "high-stakes" context. Regarding fairness, there is wide consensus that certain personal characteristics are ethically *protected* characteristics; these characteristics ought not be taken into account in high-stakes contexts—when the outcome of the decision can have a great impact on one's welfare. Loan approval decisions, hiring decisions, recidivism risk and suitability

[4] Broadly construed to include (at minimum) opportunity costs, as well as intangible/unquantifiable harms such as rights violations, insufficient or inaccurate representation, harm to social reputation, and harm to self-esteem.

for parole all seem to be areas in which we need to pay special attention to *how* judgments are made because there are *fair* and *unfair* ways of making these judgments.

Given that there are clear cases in which we do and should value fairness over efficiency, and that it seems reasonable to interpret *being treated unfairly* as a kind of harm, contexts in which judgments might be made *unfairly* should be considered a type of high-stakes context with a significant risk of harm. If so, we can incorporate considerations of fair treatment in a *general* account of contexts in which there is significant risk of harm. Unfair treatment is one among many potential harms that we risk when we employ opaque AI; I propose that we widen the category of domains in which we *might* be prohibited from using opaque AI beyond those which fit Robbins's description of "morally sensitive contexts" to include *any* context in which there is an opportunity to substantially impact the welfare or wellbeing of an individual or group. We can call these "ethically significant" contexts of use. Insofar as actions or decisions made in these areas can have significant impact on our wellbeing, special attention ought to be paid to our methods for arriving at decisions and determining our course of action in these areas. We may be ethically prohibited, for instance, from using an opaque AI model in hiring decisions because that model may exhibit unfair gender or racial bias, which has a significant impact on the welfare of those applicants. In the same way, we may be prohibited from using certain opaque AI models when deciding on actions regarding global food production: because the stability and resilience of the global food chain has a significant impact on human welfare, we may be ethically required to ensure that we have adequate understanding of the tools and processes on which we base those decisions.

The boundaries for what qualifies as an ethically significant context on my account are wide and somewhat more vague, and may cast a wider-than-expected net over contexts that qualify as "ethically significant." I believe the vagueness and breadth of this category accurately reflect the fact that our actual judgements regarding what features of the world qualify as ethically significant are notoriously difficult to codify.[5] While these judgments are sometimes unpredictable, there are also central cases on which all or nearly all can agree. Additionally, unlike Robbins, I am *not* suggesting a blanket prohibition against the use of opaque AI in all ethically significant contexts. Therefore, identifying which specific cases qualify as ethically significant will not ultimately determine whether it is ethical to employ opaque AI in such a case. Rather, identifying a context as ethically significant means that we are required to subject that case to further scrutiny before we can determine whether it is ethical to employ opaque AI.

As indicated above, a more complete account of the costs and benefits of prohibiting the use of opaque AI in certain contexts will consider contexts beyond those in which issues of bias may arise. A more inclusive (but still incomplete) account of ethically significant contexts will include contexts in which there are multiple types of opportunity cost: risk of inappropriately skewed distribution of benefits (increasing inequity) as well as risk of missed opportunity for significant benefit (especially for vulnerable populations). Recognizing these features as relevant to the ethical significance of a situation allows us to treat cases in which opaque AI may be utilized in areas such as climate science, extreme weather event prediction, public and private healthcare,

[5] See Skerker, Purves, and Jenkins (2015) on the anti-codifiability problem in robot and machine ethics.

and global food production as *ethically significant contexts*. These areas have sometimes been misidentified as areas in which ethical concerns regarding AI opacity do *not* arise, because it seems obvious that we value efficiency over transparency in such cases.[6] However, granting that we do in fact value efficiency *over* transparency in these areas does not entail that we cease to value transparency here, and it certainly does not entail that decisions and actions in these areas are ethically neutral or trivial. It would be a mistake to regard areas in which our concern for efficiency wins out over our concern for transparency as areas that are "ethically neutral", as Robbins (2019) seems to do. There are certain domains in which we value efficiency over transparency *for ethical reasons,* and to ignore this would grossly mischaracterize the domain of ethical concern. Rather, in such cases, the particular ethical concerns we have are not put in *sufficient* jeopardy by the opacity of AI to justify the missed opportunity to substantially increase human welfare, which is itself a central ethical concern.

4. Transparency as a proxy for trustworthiness (or, *If I knew what you know, I wouldn't need to* trust *you*)

An essential step towards answering the question of when, why, and to what extent we value transparency in AI is to identify the *goal* of increasing transparency. We can then ask whether that goal could be achieved by means other than transparency itself. Many have suggested that one of the main ethical goals[7] in increased AI transparency is related to *trust*: we value transparency because it serves as a proxy for the trustworthiness of the AI model.

This is similar—but in at least one sense, importantly different—to the claim that, as transparency increases, stakeholders' trust may reasonably increase as well.

Consider the domain of medical diagnostics. There is a widely-supported movement for increased transparency in the AI tools that are currently used in making medical diagnoses, and the motivation behind the movement seems to be grounded in the importance of *trust* within the medical setting and the doctor-patient relationship. *Trust* and *trustworthiness* are two distinct but related concepts, and both are essential to a successful expert/layperson relation. Whether a system or tool is *trustworthy* depends on the typical functioning of the tool—the actual predictive accuracy and reliability of the AI diagnostic tool, whether it is sufficiently robust in the face of small changes, and whether its predictions are based on a sufficiently broad and representative dataset. *Trust*, on the other hand, is a relation that holds between doctors and their diagnostic tools, or between doctors and the patients who rely on them. The presence of *trust* between doctor and patient increases the likelihood that the doctor will be able to effectively treat the patient; ideally, this improves the patient's health-related wellbeing. This trust is appropriate—when it is—in part because society has guidelines in place to ensure that a doctor's extensive training results in sound medical judgment, and a well-functioning system for verifying expertise (such as board certification and licensing).

[6] See Robbins (2019) on valuing efficiency *rather than* transparency in certain non-trivial cases.
[7] There are epistemic advantages to increasing transparency in AI models, but for the sake of this paper we are focusing solely on the ethical goals of requiring transparency in AI.

Trust is an essential feature of modern society's successful (when it is successful) division of epistemic labor. It is clearly indispensable for a successful doctor-patient relationship, and the same holds for the epistemic and ethical relationships between experts and laypeople in general. Trust is essential in the *absence* of understanding and explanation (with sufficient understanding and explanation, trust can be unnecessary). It is often thought that we trust processes that we understand, as Riberio et al. (2016) make explicit here:

> Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: if the users do not trust a model or a prediction, they will not use it. It is important to differentiate between two different (but related) definitions of trust: (1) trusting a prediction, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by *how much the human understands* a model's behaviour, *as opposed to* seeing it as a black box. (Riberio, 2016, section 1, emphasis mine)

This is a common assumption regarding the relation between trust and understanding, but it ignores an additional function and value of trust and trustworthiness. Both increased trust and increased understanding typically result in an agent's increased willingness to believe a certain decision is accurate or engage with a certain tool. But when patients trust their doctors, that trust is *not* grounded in the patients' understanding of the doctors' evidence or reasoning. This remains opaque. Patients trust their doctors because they know that, in a well-functioning social system which includes institutions dedicated to expert verification, a person would not hold the position of doctor unless they possessed the adequate expertise.

In a society that operates with a successful division of epistemic labor, trust and trustworthiness can *replace* understanding as epistemically and ethically sound grounds for belief. Laypeople believe the judgments of specialized experts because they trust those experts—*not* because they understand their reasoning—and they trust those experts because their social framework includes institutions whose role it is to verify the legitimacy of specialized experts. If the ultimate aim of increased transparency is to establish trustworthiness and build trust where appropriate, there may be other avenues available for pursuing these goals– paths that allow us to benefit from the power of opaque AI models by verifying the models' trustworthiness. Transparency itself need not be our goal.

If this is correct, then the options before us are either (1) accept that the ethical concerns which give us reason to employ opaque AI models may outweigh the benefits of transparency, and determine how to best utilize opaque AI given these epistemic limitations, or (2) refuse to employ opaque AI models in any ethically significant contexts on the grounds that the use of an opaque process is ethically impermissible in those contexts.

Given that there are enormous potential benefits that could arise from the proper use of opaque AI models in at least some of the commonly identified ethically significant domains – healthcare, climate science, the global food chain, public safety – we would need powerful ethical

reasons to support fully *eliminating* its use in these areas. The success of the expert/layperson division of epistemic labor shows us that many of our ordinary, ethically responsible, and reliable social practices already implicitly reject (2): we routinely employ opaque processes in ethically significant domains. And I will argue that there is no special reason to embrace (2) in the case of AI. If this is correct, then we are left with option (1), and the ethical question before us is no longer *whether we ought to allow opaque AI to operate in any ethically significant domains* but rather *what are the most ethical ways of harnessing opaque AI in these domains.*

5. The expert/layperson relation- a blueprint for ethical opaque AI

I have suggested that we take our successful social practice of deferring to specialized experts as a guide for developing an epistemically and ethically sound method for utilizing opaque AI models. To this end, we will need to examine when (i.e., under what conditions) it is epistemically and ethically responsible to defer to experts *rather than* relying on one's own reasoning. We also need to know *what features* make an individual a genuine expert, how, as a society, we *determine* that some individual is an expert, and what methods we use for deciding how to act when multiple experts *disagree* in their decisions. Fortunately, these questions have received substantial attention both in sociology and philosophy, under the general headings of *social epistemology* and the *epistemology of testimony*.

In what follows I will make a preliminary case for the claim that the essential features of *experts* the features that make expert opinion trustworthy, and our trust in those individuals' decisions both epistemically and ethically responsible—can be realized in AI as well. For this to be the case, the relevant features of human experts must not be *essentially human* features. Certainly, human experts have noteworthy features that AI models lack; for instance, we typically assume that human experts have a concept of the greater good and a desire to promote it. If such traits play an *indispensable* role in generating the trust and trustworthiness on which the expert/layperson relation depends, this relation will not be a viable model for the ethical use of opaque AI. As I hope to show below, the trust that exists in the expert/layperson relation is not fundamentally based on faith in the moral goodness of the expert but rather on the nature of expertise and the existence of institutions that serve to verify these experts. If these features are not uniquely human features, then, insofar as we have ethically acceptable methods of evaluating when we ought to defer to human experts in high stakes contexts, we have a potential framework for determining when it is ethically appropriate to defer to the decisions generated by opaque AI models.

In the mid 1980's, philosopher John Hardwig sparked renewed interest in the social aspects of knowledge-building by drawing attention to the myriad situations in which we are better off—rationally speaking—deferring to someone else's judgment on a particular matter rather than attempting to reason through that matter ourselves. These are situations in which the matter at hand concerns an area of highly specialized knowledge, and there are highly-trained experts who specialize in that area. In such a case, a layperson would be more rationally justified in deferring to the expert's judgment than they would in performing their own independent reasoning and

standing by the judgment at which they themselves had arrived. That is to say, a layperson has better reasons to believe an expert's judgment is correct than his or her own, even when that judgment conflicts with theirs. Assuming that the layperson is a genuine layperson, and the expert a genuine expert, Hardwig writes,

> If, then, layman B (1) has not performed the inquiry that would provide the evidence for his belief that p, (2) is not competent, and perhaps could not even become competent, to perform that inquiry, (3) is not able to assess the merits of the evidence provided by expert A's inquiry, and (4) may not even be able to understand the evidence and how it supports A's [the expert's] belief that p, can B nonetheless have good reasons to believe that A has good reasons to believe that p? I think he can. If so, should we conclude that B's belief that p is rationally justified? I think we should, acknowledging that B's belief stands on better epistemic ground than other beliefs which we would call simply irrational or nonrational. (1985, p.339)

Following Hardwig, we can say that in order for laypeople to be justified in deferring to the (opaque) reasoning of experts—rather than being rationally required to perform their own (transparent) reasoning there are (at least) three criteria that must be met[R1] .

1. The laypeople *have not*, themselves, performed the reasoning that is being left to the expert.
2. The laypeople *are not capable* of performing the reasoning that is being left to the expert (for any of several possible reasons, to be discussed below).
3. The laypeople cannot themselves 'assess the merits of the *evidence'* nor understand how the evidence supports the expert's decision. (This combines 3 and 4 in Hardwig's criteria, above).

5.1 Ruling in–and ruling out–the use of opaque AI

As will soon become apparent, even a framework intended to show where we are permitted to employ opaque AI models in ethically significant contexts will rule *against* the use of opaque AI in many of the notoriously problematic cases in which those models are already in use. Below, I will adapt Hardwig's (minimal) criteria for deference to experts to apply to AI models and briefly discuss the most readily apparent implications of interpreting each criterion in these particular ways.

1. Neither transparent models nor humans have performed the task in question on the *scale* at which the opaque AI model will be performing that task.

Explicitly specifying that the *scale* of the task is essential to properly characterizing the task itself; at the same time, drawing attention to the scale of the task makes more clear our general motivation for applying AI to any particular task at all. In broad terms, many of the same *types* of tasks that AI models are designed to perform—reviewing loan applications, evaluating job candidates, deciding how to deploy police resources, predicting effects of climate and weather events on food production—have all previously been performed by human individuals (sometimes

utilizing standard algorithms). But the *size* of the problems to which we might apply the tools of AI, the scale on which we intend for these tasks to now be performed is unprecedented, and may require more human labor-hours than we can reasonably expect from human beings within the requisite time allotted for the task.

That said, if this first criterion *must* be met for any ethically responsible application of opaque AI in an ethically significant contest, then many instances in which opaque AI has already been deployed may *not* satisfy the criteria necessary for the ethical use of opaque AI. (More will be said about this when we discuss guideline (B) in the following section.)

2. Transparent models are *not practically capable* of performing the task that the opaque AI model is intended to perform.

Whether this criterion is met will in part depend on the state of AI technology and the actual skillsets of AI researchers at the time the decision is being made. Rudin (2019) points to this aspect of the problem when she writes,

> Black box models seem to uncover 'hidden patterns'. The fact that many scientists have difficulty constructing interpretable models may be fueling the belief that black boxes have the ability to uncover subtle hidden patterns in the data about which the user was not previously aware. A transparent model *may* be able to uncover these same patterns. If the pattern in the data was important enough that a black box model could leverage it to obtain better predictions, an interpretable model might also locate the same pattern and use it.
>
> Again, *this depends on the ML researcher's ability to create accurate yet interpretable models.* The researcher needs to create a model that has the capability of uncovering the types of pattern that the user would find interpretable, but also the model needs to be flexible enough to fit the data accurately. *This, and the optimization challenges discussed above, are where the difficulty lies with constructing interpretable models.* (2019, p201, emphasis mine)

If equally proficient transparent models[8] already exist or could realistically be developed within the requisite timeframe (where 'equally proficient' takes into account the *speed* required to perform the task effectively as well as the *scale* of the task), the additional value conferred by their transparency may make them ethically preferable to an opaque model. Though Rudin is optimistic regarding the potential of transparent (in this case, interpretable) models to perform as well as opaque models, this is by no means guaranteed. As she acknowledges, "This problem is compounded by the fact that *researchers are now trained in deep learning, but not in interpretable ML*..." and "It could be possible that there are application domains where a complete black box is

[8] While "opaque" has a standard meaning in the literature on this topic, "transparent" has several common meanings when used in the context of AI models. A satisfactorily transparent AI model might be an interpretable model, or an explicable model, or it may be comprehensible to the relevant practitioner or stakeholder, etc. A thorough account of how "transparency" has been interpreted in the literature on AI regulations is beyond the scope of this discussion, but see Lipton, 2016; Miller, 2017; Mittelstadt, et al. 2019; Molnar, 2019; Riberio, 2016; Rudin, 2019.

required for a high stakes decision," though she notes that, "As of yet, I have not encountered such an application" (2019, p207).

3. We are unable to satisfactorily explain the AI model within a reasonable amount of time given the urgency of the task in question.

An explanation of an AI model would allow us to "assess the merits" of the evidence on which the model is basing its decision and "understand… how [the evidence] supports" that decision. The third criterion roughly specifies that in order for us to sacrifice transparency for the benefits gained by employing opaque AI in a particular ethically significant context, that opacity must be a result of our genuine *inability* to explain the operations of the AI model, rather than an *unwillingness* to deploy sufficient resources to the task. (Note that this issue will only arise when there is a question of *irresponsibly* employing opaque AI– the context itself must be *ethically significant* for ethical concerns to compete with the value of the efficiency or accuracy gained by utilizing opaque AI models.)

In addition to this cursory description of when it would be reasonable for a layperson to defer to the judgment of an expert, Hardwig also provides a rough approximation of the personal features that make an individual an expert. Briefly, an expert must have engaged in "inquiry that has been sustained, prolonged, and systematic" (1989, p. 338). Though we would need to determine what features of an AI model would make its "inquiry" into a specific domain suitably "sustained, prolonged, and systematic," this criterion seems to pose no special difficulty for AI. And given that these models fundamentally function by discovering and attuning themselves to patterns in data, these data-processing operations should satisfy all relevant features of an "inquiry."

5.2 The social institutions/practices underwriting our successful deference to experts (and how they might be replicated in the case of AI)

So far I have proposed a set of fundamental criteria that would need to be met in order for an individual—or an AI model—to qualify as an expert, as well as conditions under which may it be epistemically and ethically responsible to defer to the judgments of a human or artificial "expert". In this section we will consider preliminary ideas regarding how we might *determine* whether some opaque AI model should be considered an expert in this sense. An opaque model may possess the requisite features for "expertise" in a certain area, but the opacity of that model will make it challenging for us to *know* whether the model has satisfied the appropriate criteria. In addition, I will make preliminary suggestions for how we might deal with morally weighty cases in which (just as with human experts) multiple opaque AI models *disagree* in their predictions or decisions.

In the familiar cases of human experts, the answers to both of these questions rely, in part, on the existence of a *larger network* of experts in place in addition to the individual (potential) expert in question. In areas of technical specialization (academic research, professions such as journalism and law, etc.) we commonly find established institutions and professional organizations that grant degrees, credentials, or otherwise certify that the individual in question does in fact

qualify as an expert. These organizations are typically composed of individuals who themselves possess certain types of relevant expertise. When cases arise in which a purported 'expert' fails to meet the standards set by the certifying bodies in their fields, we rely on these institutions to revoke that individual's credentials. Lawyers can be disbarred, doctors can lose their license to practice medicine, journalists can lose their press credentials, and so on. Ideally, this process serves to inform the general public that these individuals are not, in fact, genuine experts in their supposed fields. These institutions allow laypeople to know which individuals are experts in which fields, and responsibly defer to their judgments, even though *exactly what makes* that individual an expert in that field is beyond the understanding of the layperson.

The presence of multiple experts within a single field is not only essential to our ability to know *which individuals* are experts (since we, as laypeople, cannot evaluate their expertise for ourselves); the fact that large numbers of independent experts regularly converge in their opinions give us an imperfect but reliable indication that these judgments are correct, as well as a means of determining how to act when experts *disagree*. If a significant majority of genuine experts converge in their opinion on a particular issue, and a small number of experts disagree, it will be *reasonable* for the layperson to accept the opinion of the majority.

Adapting our methods for certifying experts and handling expert disagreements such that we can apply them to opaque AI models presents more of a challenge than adapting the criteria for expertise itself or for responsibly deferring to experts. The relationship between laypeople and experts in modern society has a long history, and the trustworthiness of these credentialing institutions is born out only by society's *repeated knowledge-building success over time.* Our engagement with opaque AI models has both a short and checkered past. We do not have the convenience of a lengthy history—on a human timescale—to indicate which methods for certifying the expert-status of an opaque AI model will prove to be trustworthy, and which methods are likely to fail.

Because of the importance that *time* plays in revealing the reliability of expert decisions, of our method for verifying individuals as genuine experts, and of our division of epistemic labor in general, whatever way in which we choose to adapt this feature to create an analogous method for revealing the trustworthiness of opaque AI models will be highly speculative. There are no obvious candidates for artificial analogues of the passage of time. With that in mind, one possible option would be to treat the notion of an epoch in artificial neural networks as a stand-in for the ordinary passage of time. Rather than thinking of the history of AI models on a human timescale, it may be more appropriate to frame the notion of "an adequate length of time" on which to judge the reliability of an AI model to reflect an AI timescale. So whereas ANNs and other deep learning models may have emerged 10 years ago on a human timescale, a massive number of epochs for those models has passed within this span. (Determining the optimal number of epochs for training a neural network is currently considered something of an art in machine learning.)

There are a growing number of organizations dedicated to developing something akin to a "credentialing processes" for AI. The Institute of Electrical and Electronics Engineers (IEEE) continuously updates its standards for the development and use of AI. The International

Organization for Standardization (ISO) and The International Electrotechnical Commission **(**IEC) both work to develop standards that aim to make AI more "resilient, reliable, accurate, and secure". And the European Commission's 2021 proposal for Regulation on Artificial Intelligence includes a legal framework by which to judge the risk of AI. The UK Institute for Ethical AI and Machine Learning, the Global Partnership on AI (GPAI), and the OECD AI Policy Observatory all support projects and policy aimed at increasing trustworthiness in AI. What form a successful credentialing process will eventually take, and to what extent these certification systems are already in place, is a question to be addressed elsewhere. But if we are interested in developing an approval process that could certify AI models and approve their use in particular contexts *while allowing* these models to remain *opaque*, we might make progress on this issue by continuing research into the relevant features of familiar and successful practices of certifying human experts.

The final feature of the expert-layperson relationship that we will address here—our methods for dealing with cases of expert disagreement—is simple to adapt in theory (though perhaps less so in practice). Our successful division of epistemic labor crucially depends on the existence of multiple independently trained experts in a single field, addressing the same issue and converging on the same opinion through a variety of independent methods. At the present moment, it is unclear whether there exists a sufficient number—and variety—of AI models that could perform the same ethically significant task (whatever this task may be) in order to deal with disagreement in an analogous way. But there may be no *better* way to establish the requisite level of trustworthiness[9] [1] of an opaque AI model than developing multiple, independent, opaque models, operating with distinct architecture and trained on distinct (but appropriately similar) data sets, and finding that they converge on the same decision. Given that opaque AI will be an ever-present ethical issue, developing multiple models to perform the same ethically significant task may be well worth the investment.

6. Preliminary guidelines for the ethical use of opaque AI

Given that I claim it may be ethically permissible (perhaps required) to use opaque AI models in certain ethically significant contexts, this section provides a plausible decision procedure for evaluating whether a particular context is one in which we could ethically employ opaque AI. I suggest three general questions that should be addressed in the process of making such a decision. The first two questions are as follows:

(A) Is the context in question an ethically significant context?

(B) Could the task at issue be performed *equally well* by a transparent process?

(C) Are the benefits of successfully performing this task greater than both i) the cost of potentially failing at this task (whatever constitutes "failure" in this case) and ii) the cost of not performing this task at all?

(A) Is the context in question an ethically significant context?

[9] to whatever extent is required such that it would be ethically responsible to utilize that opaque model in the particular ethically significant context in question

If we can be reasonably certain that the answer to (A) is "no," then the ethical concerns surrounding the use of opaque AI do not arise in this situation, and we are at liberty to use opaque AI for the task at issue. Note that the triviality or ethical significance of a context will be decided according to a broad and diverse set of standards, some of which may involve apparently objective and quantifiable measures (for example, the potential consequences of utilizing some proposed AI model in the global food supply chain) and some of which may involve standards that will vary relative to a cultural or social context (the impact of utilizing some proposed AI on the representation of a particular socially marginalized or vulnerable group). Note also that the ethical significance of a context will be a matter of degree, depending on the gravity of the particular situation(s) involved.

If the answer to (A) is "yes," then we need to address question (B). Could the task at issue be performed *equally well* by a transparent process (whether human or AI)? This question will be familiar from the criteria for rationally deferring to experts in general. The additional benefits that arise from *transparency in how decisions are made* in all ethically significant contexts may outweigh whatever benefits the opaque model may provide. Here, it is important to note that "performing a task equally as well" will include—at minimum—issues of equity and fairness *in addition to* efficiency and accuracy. As noted in section 3, we cannot entirely ignore the harms of opportunity costs for the sake of eliminating bias, *especially* when those costs are borne by marginalized and vulnerable populations.

To permit the use of opaque AI in an ethically significant context, the answer to question (A) must be yes, and the answer to question (B) must be no. If so, then it may be ethically permissible to utilize opaque AI, subject to further consideration, such as those raised in question (C). Are the benefits of successfully performing this task greater than both i) the cost of potentially failing at this task (whatever constitutes "failure" in this case) and ii) the cost of not performing this task at all?

If there are ethically significant cases in which all three bars are met, then there are non-trivial cases in which we would be permitted—perhaps required—to utilize opaque AI. And given that meeting all three bars requires that the opaque model in question be reliable and trustworthy, we will need a framework for evaluating the reliability and trustworthiness of opaque AI models. I hope to have made a preliminary case for looking to our successful social practice of deferring to experts in ethically significant domains for a blueprint of how to responsibly employ opaque AI in such a case.

Conclusion

I acknowledge that, even as guidelines go, those given above are considerably vague. I view this vagueness as appropriate, and—practically speaking—ineliminable. Here, we are concerned with developing rules for ethical action in the use of AI, and as Aristotle said, we should only look for precision in each class of things just so far as the nature of the subject admits. Any rule, no matter how precise, requires interpretation when applied to a particular case. And when the interpretation of those rules involves disentangling and weighing competing moral values, it is

the process of interpretation itself—and not the rule—that will be doing the lion's share of the work. So I would suggest that insofar as these guidelines are vague, their vagueness is appropriate to the subject at hand. Deciding whether a task could be performed *equally well* by some satisfactorily transparent (human or algorithmic) decision-making process will involve weighing competing values, and the relative strength of those competing values will depend on the ethical inclinations of the individuals performing the evaluation. There is no standard, universally applicable measure for assigning weight to these values; each case will need to be evaluated individually, and an argument will need to be made for weighting any of these values more strongly than the others. The same is true for deciding whether the benefits of success are worth the potential costs of failure. Human judgment cannot be entirely removed from decision-making in ethically significant domains, no matter how trustworthy the AI model involved. At minimum, humans must still be in-the-loop in order to 1) make case-specific value-judgments, and 2) make cost/benefit assessments in cases where the costs and benefits are not fully commensurable. And given that we are discussing opaque AI models, humans will need to be in-the-loop to monitor for potential instances of biased outcomes. The threat of bias will remain, whether or not the cost of that potential bias is outweighed by the potential benefits of a successful outcome.

These guidelines are not intended to serve as a complete checklist for the ethical use of opaque AI. They merely offer one plausible set of rules for evaluating whether some instance is an instance in which we should consider, or refuse, to employ an opaque AI model to a task. If we should, we might then look to the blueprint provided by the expert/layperson division of epistemic labor to see how to do so *well*. In addition, the overview of the expert/layperson relation given above is not intended to fully capture the robust and complex features of this social epistemic practice. Whether this overview accurately represents the fundamental features of this relationship is separate from the question of whether the expert/layperson relation itself—and the institutions that support it—can provide us with a general framework for developing an ethical approach to harnessing the power of opaque AI, as I believe it can.

Bibliography

Ahmed, M. (2018). Aided by Palantir, the LAPD uses predictive policing to monitor specific people and neighborhoods. *The Intercept*. https://theintercept.com/2018/05/11/predictive-policing-surveillance-los-angeles/.

Barry-Jester, A., Casselman, B., & Goldstein, D. (2015). The new science of sentencing. *The Marshall Project*. https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

Barocas, Solon. (2018) "Accounting for Artificial Intelligence: Rules, Reasons, Rationales." *Human Rights, Ethics, and Artificial Intelligence*, 30 Nov. Harvard Kennedy School Carr Center for Human Rights Policy. Lecture.

Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, 13(1), 94–115. https://doi.org/10.1111/jels.12098.

de Bruijne, M. (2016). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, 33, 94–97. https://doi.org/10.1016/j.media.2016.06.032.

Chen, C. et al. This looks like that: deep learning for interpretable image recognition. Preprint at https://arxiv.org/abs/1806.10574 (2018).

Dhar, J., & Ranganathan, A. (2015). Machine learning capabilities in medical diagnosis applications: Computational results for hepatitis disease. *International Journal of Biomedical Engineering and Technology,* 17(4), 330–340. https://doi.org/10.1504/IJBET .2015.06939 8.

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *RadioGraphics*, 37(2), 505–515. https://doi.org/10.1148/rg.2017160130.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s1102 3-018-9482-5.

Future of Life Institute (2017) Asilomar AI Principles. https://futureoflife.org/ai-principles/.

Goldman, A.I. (2001). Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research*, 63(1): 85–110.

_____. (2014). "Social Process Reliabilism: Solving Justification Problems in Collective Epistemology", in Lackey 2014: 11–41. doi:10.1093/acprof:oso/9780199665792.003.0002

Günther, Mario & Kasirzadeh, Atoosa. (2022). Algorithmic and human decision making: for a double standard of transparency. AI & SOCIETY. 37. 10.1007/s00146-021-01200-5.

Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy*, 82(7): 335-349.

Lackey, J. (2016). What Is Justified Group Belief? *Philosophical Review*, 125(3): 341–396. doi:10.1215/00318108-3516946

Li, O., Liu, H., Chen, C. & Rudin, C. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In Proceedings of AAAI Conference on Artificial Intelligence 3530–3537 (AAAI, 2018).

Lipton, Z. C., (2016). The mythos of model interpretability. In: ICML Workshop on Human *Interpretability* in Machine Learning, vol. 2017, pp. 96–100, 24

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report* 49, no. 1 (2019): 15-21. DOI:10.1002/hast.973

Miller, Tim. Explanation in Artificial intelligence: Insights from the social sciences (2017) arXiv

Mittelstadt, B., Russell, C. & Wachter, S. Explaining explanations in AI. In Proceedings of Fairness, Accountability, and Transparency (FAT*) (ACM, 2019).

Molnar, C. (2019). Interpretable Machine Learning.

Morrell, A. (2018). Citigroup has inked a deal with an AI-powered fintech to help flag suspicious payments and safeguard a $4 trillion daily operation. Business Insider. https://www.businessinsider.com/citi-has-inked-a-deal-with-an-aipowered-fintech-feedzai-2018-12.

Nadella, S. (2016). Microsoft's CEO explores how humans and A.I. Can solve society's challenges— together. *Slate*. https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societyschallenges.html.

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

Public Voice, The. (2018). Universal Guidelines for Artificial Intelligence. https://thepublicvoice.org/ai-universal-guidelines/

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. (2016) "why should I trust you?": explaining the predictions of any classifier. KDD.

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds & Machines* 29, 495–514. doi:10.1007/s11023-019-09509-3

Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead Nature Machine Intelligence | 206 VOL 1 | MAY 2019 | 206–215 |

Skerker, M., Purves, D., and Jenkins, R. (2015) Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice.* Vol. 18, No. 4, Special Issue: BSET-2014 pp. 851-872

Vincent, J. (2018). AI that detects cardiac arrests during emergency calls will be tested across Europe this summer. *The Verge.* https://www.theve rge.com/2018/4/25/17278 994/ai-cardiac-arrest-corti-emergency-call-response.

Whittaker, M. et al. (2018) *AI Now Report*.

Zerilli, J., Knott, A., Maclaurin, J. *et al.* (2019) Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?. *Philos. Technol.* 32, 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zerilli, J. (2022) Explaining Machine Learning Decisions, *Philosophy of Science* 89 (1):1-19