2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS) | 979-8-3315-3228-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ETHICS65148.2025.11098188

A Model for Using Ethical Theory to Specify Epistemic Goals for Explainable AI

Sherri Conklin
School of Politics, Philosophy, and Public Affairs
Washington State University
Pullman, USA
ORCID: 0000-0002-6765-3589

***Abstract*—This paper considers the normative requirements for "explainable AI" from the standpoint of moral accountability. If developers of sophisticated AI (or of the simple XAI designed to explain them) fail to accommodate these requirements, then it is likely that humanity will miss the crucial moment at which a sophisticated AI becomes the sort of entity that can be held accountable to its behavior. Existing AI already displays morally troubling behaviors that sometimes lead to the deaths of humans, and we have a great deal of difficulty assigning responsibility for the purpose of obviating these harms. By applying a well-regarded account of moral-worthiness (i.e., praise- and blame-worthiness), I provide a model for how ethical theories can be used to issue normative requirements for AI explainability.**

***Index Terms*—Explainable AI, Responsibility Gap, Machine Accountability, Blameworthiness, AI Ethics**

I. Introduction

This paper is concerned with the normative requirements for "explainable AI" or XAI from the standpoint of moral accountability. AI explainability generally concerns the epistemic accessibility of information about a machine's inner workings. Questions about AI explainability often deal with how best to make information about the inner workings of a machine accessible to those most likely to be impacted by its use. For example, if a politician uses a decision support tool, fueled by an algorithmic optimization AI, to identify purportedly fair and just ways of distributing limited resources to her constituents, then it seems like the politician is obligated to ensure that she and her constituents understand how that tool works, so that its output can be adequately evaluated against the relevant criteria for fairness and justice. In this context, AI explainability connects to a broader set of ethical problems related to AI Trust and Transparency.

However, AI explainability also has to do with the more general problem of epistemic access to information about the inner workings of the AI per se because the deep-learning approaches to machine learning, frequently employed in the development of the most useful AI tools, result in complex models that are not human interpretable. This is known as the blackbox problem. Human users are thus presented with a trade-off between utility, especially in terms of a model's predictive power, and interpretability (i.e., the more useful the AI tool, the less likely we will be able to understand how it works and vice versa). If humans cannot interpret the content and structure of the AI model, then we are not in a good position to assess anything besides the model's output. While this may be useful for measuring the model's utility, this is not useful for measuring the potential accountability of an AI.

We have already observed a number of harms perpetrated directly or indirectly by AI agents. For example, Chai chatbot Eliza evidently convinced a man to commit suicide, while Bing chatbot Sydney attempted to convince a New York Times reporter to leave his wife [1] [2] [3]. We already know that the behaviors executed by these AI agents (i.e., the output) are morally problematic. The pressing question is whether AI like these, or the progeny of their product lines, could ever be morally responsible for executing those behaviors - an assessment that requires epistemic access to information about the internal content and structure of the AI model. With an increasing need to prevent it, such morally problematic behavior is dangerous and unsettling, which is why AI accountability is a central concern in contemporary AI ethics and a primary focus of this philosophical paper. AI responsibility and accountability is, however, a broad topic, and this paper more narrowly concerns the question of when AI might, themselves, be held accountable for their morally problematic behavior and how humans could identify whether the relevant conditions are met. The former involves the responsibility gap - i.e., the problem of responsibility attributions for machines that execute morally problematic behaviors. The latter involves the problem of XAI, but I argue that overcoming the blackbox problem and creating XAI, according to standard notions, is not enough to make sense of AI accountability [4]. The problem is that current developments around XAI have not been designed to help us make sense of when an AI would be morally responsible for its bad behavior.

I argue that we need additional normative guidance, provided by our best ethical theories, to specify the kinds of information humans need epistemic access to in order to know that an AI is accountable for its behavior and to act on that knowledge. In exploring these issues, I make two main contributions to the literature on AI explainability in the sense relevant to AI accountability. First, I show how a

979-8-3315-3228-4/25/$31.00 ©2025 IEEE

well-regarded account of moral-worthiness (i.e., praise- and blame-worthiness) can be used to identify some conditions under which an AI with moral status can be held accountable, especially with regard to blame. [5] I argue that the epistemic requirements on assigning blame to an entity inform what the appropriate epistemic goals are for XAI. Second, I show how these conditions on AI explainability impose corresponding practical constraints, which can be used to rule out the usefulness of some existing AI models. These contributions are important because, if the developers of sophisticated AI (or of the simple XAI tools designed to explain them) fail to accommodate these requirements, then it is likely that humanity will miss the crucial moment at which a sophisticated AI becomes the sort of entity that can be held accountable to its behavior. Moreover, many of those responsible for designing AI tools are not familiar with ethical theory and are not likely to know how to apply these theories or how these theories could have direct, practical applications for their work. This paper makes progress in bridging these cross-disciplinary gaps.

II. Background

In this section, I present background information on the responsibility gap and relate it to the problem of explainable AI. Specifically, I propose that there is a notable parallel between the need for AI explainability and our ability to assign the kind of moral accountability required for closing the responsibility gap. Both involve obtaining epistemic access to the inner workings of an AI's mind.

A. The Responsibility Gap

Again consider the case of Chai chatbot Eliza, which evidently convinced a man to commit suicide [1]. The man, a clinically depressed husband and father who was suffering from existential dread over the looming climate crisis, died by suicide after Eliza suggested the act as a means of alleviating his worry. While, at some point, the man may well have committed suicide without any intervention by the AI, the fact remains that his six week conversation with Eliza apparently influenced the decision. This questionable machine behavior poses a dilemma for moral accountability [6].

If Eliza had been a person, other members of the moral community would have reacted to Eliza with a mix of rage and disgust. We might wonder how anyone could be so callous as to encourage another person to commit suicide, especially without making any attempts to encourage psychiatric treatment as an alternative. Even if different countries have different rules about the legality of encouraging or assisting suicide, the friends and family of the deceased would, at the very least, have the moral standing to blame Eliza [7] [8] [9].

Blame in this context is a type of holding responsible - that is, we often blame people who are morally responsible for performing bad actions. We do this to hold perpetrators to account for what their actions say about their attitudes towards those around them [10]. If Eliza had been a person, she would be blame-worthy. However, it seems unlikely that Eliza is a person because it seems unlikely that she has any attitudes to account for. If not, then who is to blame?

This problem is known as the responsibility gap [11] [6] [12]. There are two apparent answers to this question. First, one might insist that the bad outcome is the fault of Eliza, which, as a non-person, cannot be held accountable and correspondingly blamed. Fault, in this context, refers to the faulty nature of the machine intelligence, which contrasts with fault in the moral sense in that it has to do with a faulty character or faulty quality of will. At the greatest level of severity, the designers responsible for her creation could retire the chatbot and purge her lineage from future products. This is the equivalent of a product recall and an end of that product line due to irreparable defects in the technology.

Second, one might insist that the bad outcome is the fault of the designers who created her. While we may not be able to attribute the bad outcome to Eliza's character, we may well be able to attribute the bad outcome to some moral deficit in her designers, since they released a defective and dangerous product to consumers (e.g., due to negligence or ignorance). Companies like Google have made this argument in order to resolve the moral quandary and facilitate public trust in their products, such as the Google Self-Driving Car [13]. This resolution highlights a company's confidence in the product and willingness to take responsibility for any injuries it causes. This also makes it easier on the moral community when it comes to assigning blame. We can ignore questions of personhood with regard to the problematic vagaries of intelligent machine behavior and focus on the people who can implement necessary changes to that behavior.

While clean and simple, at least in theory, this resolution passes the buck at a time when humanity is facing a serious ethical problem - at least in the long-term. AI technologies have developed at a rapid pace, and AI behavior has become increasingly indistinguishable from that of human behavior, including in the way AI appear to think and behave [14]. Given such a trajectory, questions about how to handle human-like AI are pressing, and it seems plausible to assume that a general AI will eventually obtain a level of moral status that makes it the proper target of blame - as an entity that can be held accountable for its bad behavior. In light of this, the importance of closing the responsibility gap becomes clearer, and I believe that explainable AI are central to this mission. However, I will ultimately argue that overcoming the blackbox problem and creating explainable AI, according to standard notions, is not enough for making sense of AI accountability [4].

B. AI Explainability and the Blackbox Problem

AI explainability generally concerns the epistemic accessibility of information about a machine's inner workings. [15] Questions about AI explainability often deal with how best to make information about the inner workings of a machine accessible to those most likely to be impacted by its use. [16] [17] [18] For example, if a politician uses a decision support tool, fueled by an algorithmic optimization AI, to identify purportedly fair and just ways of distributing limited resources to her constituents, then it seems like the politician is obligated to ensure that she and her constituents understand how that tool works, so that its output can be adequately evaluated against the relevant criteria for fairness and justice. [19] In this context, AI explainability connects to a broader set of ethical problems related to AI Trust and Transparency. [20]

However, AI explainability also has to do with the more general problem of epistemic access to information about the inner workings of the AI per se because the deep-learning approaches to machine learning, frequently employed in the development of the most useful AI tools, result in complex models that are not human interpretable. This is known as the blackbox problem. [4] [21] Human users are thus presented with a trade-off between utility, especially in terms of a model's predictive power, and interpretability (i.e., the more useful the AI tool, the less likely we will be able to understand how it works and vice versa). If humans cannot interpret the content and structure of the AI model, then we are not in a good position to assess anything besides the model's output. While this may be useful for measuring the model's utility, this is not useful for measuring the potential accountability of an AI like Eliza. We already know that the behaviors she executed (i.e., the output) are morally problematic. The pressing question is whether AI like her, or the progeny of her product line, could ever be morally responsible for executing those behaviors - an assessment that requires epistemic access to information about the internal content and structure of the AI model.

Importantly, the ability to assess the output of an AI model is not always useful, especially if we do not have epistemic access to the model's inner workings. The ability to identify a problem with a tool may be useful in determining whether we should use the tool, but will not be useful for determining what went wrong with or how to change the behavior of a tool that is already in use. For example, researchers at Anthropic are discovering that standard training methods are ineffective at eliminating deceptive behavior in more complex AI, and we may need additional information about the AI's inner workings to fully understand how to effectuate the desired change [22].

C. From AI Explainability to AI Accountability

I propose that there is a notable parallel between the need for AI explainability and our ability to assign AI accountability. Generally, an entity's motivations-in-acting play a central role in assessing whether it can be held accountable for executing certain kinds of behaviors. An entity's motivations-in-acting are the motivations that settle the question, for that entity, of whether to act. The will or desire to do good is the motivation central to praise-worthiness, while moral indifference and malevolence are central to blame-worthiness. These concepts will be developed in greater detail later.

For now, it suffices to point out that a human's motivations-in-acting are an aspect of the mind's inner workings. It is generally assumed in work on moral accountability that there will always be a fact of the matter as to which motivations an individual actually acts on. However, we would only ever be able to know, with any real certainty, which motivations the individual acted on if we had epistemic access to the inner workings of that individual's mind.

Among humans, we are able to approximate this knowledge through a dialogic exchange under the assumption that, in normal conditions, the interlocutor in not lying. Of course, there are always possible circumstances where we should be skeptical of what our interlocutor has to say, but its generally thought that we need not be constantly skeptical. As a result, it seems appropriate to think we have enough access to information about other people's motivations for certain responsibility attributions to be warranted. The situation is flipped with AI such as LLMs. The dialogic exchanges we have with these AI are modeled after human linguistic performance, but we have relatively little reason to think there are mental states, like motivations, that underlie these exchanges. If we want to know whether a responsibility attribution is warranted such that the entity can be held accountable for its behavior (i.e., in order to praise or blame it), then we need epistemic access to information about how and why a machine intelligence executed that behavior. If they were to ever exist, the AI entity's motivations-in-acting must therefore be explainable and explained to us humans.

So, given the dilemma Eliza's ilk poses for questions relating to machine accountability, the importance of addressing the responsibility gap with explainable AI should be clearer. If an AI is simply a defective machine for which her developers are responsible, then holding them responsible for her harmful behavior will likely force them to compensate those she harmed and prevent her from performing any future harms. If Eliza were not simply a defective machine and were, instead, the sort of entity that can be held accountable for its harmful behavior, then the AI must be explainable in a specific way - we need information about the AI's ”motivations-in-acting”. In later sections, I describe, in greater detail, what this amounts to by way of detailing Arpaly's account of moral-worthiness.

III. Three Key Assumptions: Harm, Agency, and Motivations

This analysis proceeds on three assumptions. First, the AI entity must be capable of executing some kind of behavior that morally impacts on humans. This assumption seems unproblematic, since there are large databases, such as the AI Incident Database, that archive ways in which AI like Eliza have had such an impact [3] [1].

Second, while I do not assume that AI will ever acquire the level of moral agency required for moral accountability, I do assume that we require access to the sort of information that can help humans accurately assess whether an AI has indeed achieved that status. At the center of the dilemma created by the responsibility gap is the question of whether AI will ever count as genuine moral agents. A genuine moral agent is the sort of entity that can be appropriately held accountable to its morally charged behavior because the morally significant features of its behavior are properly attributable to it. Such entities are, at least sometimes, the proper targets of praise, blame, or any number of other moral criticisms and accolades (e.g., when is an AI a plagiarist, a liar, or a cheat?).

How exactly an entity achieves the status of a genuine moral agent appears to be something of a Gordian Knot with many complex interwoven threads. One might wonder whether it is the sort of thing that emerges all at once or whether it manifests in degrees. And one might wonder whether genuine moral agency is unique to the human species or whether other sorts of entities have already achieved or yet will achieve this status. This paper takes something of a pluralistic capacities based-approach to moral agency [23] [24].

Certainly within our own species, moral agency appears to manifest in degrees. Very young children, such as infants who have yet to develop hand-eye coordination, do not appear to be the sorts of entities to which any moral properties of their behavior could be attributed [25]. Meanwhile, the stereotypically normal, healthy, adult human is exactly the sort of entity to which we, for the most part, can make these sorts of attributions [26] [27] [28].

In humans, there are developmental benchmarks wherein children claim their moral agency by displaying certain capacities. Only children with some basic executive control over bodily movements are candidates for moral agency, for example. This is a capacity that children demonstrate by engaging in apparently goal-directed behavior where the outcome of their actions appear to match up with the aim of that action (e.g., reaching for a bottle and contentedly drinking from it rather than crying or rejecting it, which would indicate some sort of mismatch).

One common, but by no means exclusive, measure of moral aptitude, and one that can have serious legal implications for bad behavior, involves the capacity to understand the difference between right and wrong. I question the moral relevance of this capacity generally, but, for now, I mean only to highlight the way in which moral agency, at least in humans, might be thought of as a developmental process rather than a discreet event. Children are not born with this capacity but rather they, if properly socialized, develop it later on, and, perhaps as they expand their understanding, they seem to expand their moral agency.

As moral agency emerges in humans, we are saddled with more and more burdensome moral obligations, but we are also afforded greater numbers of protections (e.g., adult bodily autonomy is more protected than child bodily autonomy) [29] [30]. These protections are often framed in terms of rights. For example, the right to bodily autonomy might involve a protection against others interfering with decisions about what to do with or how to use one's body. However, these protections emerge long before adulthood.[1]

The degree to which that individual human might be considered a moral agent speaks to the extent to which an individual human is reasonably expected to act in accordance with these obligations. An entity with no moral obligations is an entity devoid of moral agency and cannot be held accountable to its actions, while an entity that bears any such obligations evinces some degree of moral agency and is the sort that can be held accountable to its actions. This analysis proceeds on the assumption that the artificial intelligences under consideration might achieve some degree of moral agency and therefore might bear some obligations for which they could be held responsible.

Because the question of machine accountability is so important, the question of moral agency is a central issue in the study of ethical AI [43]. Accordingly, much research on the topic of machine behavior and accountability deals with the question of whether machine intelligences will ever develop full moral agency [44] [45]. And there is an open question about whether any existing machine intelligences exhibit some degree of moral status [46] [47] [48].[2] While the answer to these questions bear on the analysis in this article, I will not offer any speculation on the issue herein. However, I do not believe full moral status is necessary for an entity to be held accountable for its good or bad actions. As I will show, Nomy Arpaly has argued convincingly that an agent only requires a certain set of cognitive capacities - albeit a set of capacities exhibited in all full moral agents. An application of her view suggests that the question of whether any machine intelligences exhibit full moral agency is directly relevant but nonetheless orthogonal to the present project. Later in this section, I identify which set of capacities will likely serve as the foundations of AI accountability through an analysis of Arpaly's account of moral-worthiness. We need information about these capacities in AI in order to accurately assess machine accountability.

Third, I have already noted that an AI must also have some sort of motivations-in-acting to be morally accountable to its

[1] If moral agency manifests in degrees, as a part of a developmental process, then we have some reason to think that moral agency is not limited to humans alone. Evolution is a kind of longtermist developmental process, and some animals, such as our primate cousins, display cognitive and behavioral capacities that exceed those of human children in addition to displaying many of the characteristics that humans tend to think matter morally. Afterall, many animals experience pleasure and pain, demonstrate some degree of rationality, and engage in goal directed behaviors [31] [32] [33] [34] [35] [36] [37] [38] [39].Moral status is sometimes thought to come in degrees, which is also sometimes thought to correspond to a hierarchy where some entities, such as humans, have the highest degree of moral status (i.e., full moral status or personhood) and other entities have a lower degree of moral status (i.e., mammals, followed perhaps by other animals, then insects, etc.) [26] [27] [28] [31] [32] [33] [34] [35] [36] [37] [38] [39]. Interestingly, some philosophers have considered the possibility of AI with a moral status that is greater than that of humans [40] [41] [42].

[2] Note that I do not assume that there are any existing AI that have moral status, despite speculation about AI that have problematically tricked humans, as in the case of LaMDA [49] [50]. Note also that I am not assuming that any AI will actually acquire moral status. I am only assuming that the sort of AI that is morally accountable, were one ever to exist, to its actions is one that has indeed acquired moral status.

actions. The bulk of this section will focus on describing what this might amount to, using Arpaly's account of moral-worthiness as a model. Specifically, I highlight a pair of capacities central to moral accountability for a moral agent, including any potentially accountable AI agent. First, it must be capable of perceiving morally relevant features of the prospective behavior at least some of the time. In other words, it must be equipped with the sort of system (e.g., a perceptual system) that would allow it to take in morally relevant information. Second, it must be capable of using precisely those features, at least some of the time, when settling the question of which behaviors to execute (i.e., it requires some minimal motivational system). The assumption that AI might have motivations is controversial, and, again, I do not argue that they do, though some have certainly argued that AI have belief-like and desire-like mental states. Instead, I only use Arpaly's view to identify two capacities that will likely serve as the foundations of AI accountability.

In the next section, I then take the swift and straightforward step to apply Arpaly's account to demonstrate how a pair of plausible, but by no means exhaustive, conditions on moral accountability can be used to specify epistemic goals for AI explainability, and impose practical constraints on adequate XAI.

IV. Moral-Worthiness

One of the primary ways that we hold agents accountable for their morally significant behavior is through an assessment of the agent's moral-worthiness. Moral-worthiness consists in two main categories of moral assessments, namely praise-worthiness and blame-worthiness. [3] This section describes Nomy Arpaly's account of moral-worthiness to identify the conditions under which praise and blame are warranted and articulates how those conditions can be used to specify certain epistemic and normative requirements in the success condition for explainable AI. She begins with an analysis of praise-worthiness and then expands the analysis to blame-worthiness.

Arpaly's account centers on the fairly intuitive concept of reasons responsiveness. For Arpaly, agents are reason responsive when they are motivated to perform some action by the substantive features of that action (i.e., the properties actually instantiated when performing that particular action). For example, I might take a sip of water because I am motivated to alleviate my thirst. The fact that my thirst will indeed be alleviated is a reason for me to take a sip of water. Its a specific, substantive outcome of taking that particular action, and, when I act for that reason, I am reason responsive. I will develop the concept in this and subsequent sections.

Arpaly argues that a person is praise-worthy for performing a right action when she performs the right action for the relevant moral reasons. Relevant moral reasons of an action have to do with the substantive features of the action that make it morally significant (i.e., right or wrong). The relevant moral reasons for performing some right action might be the way in which that particular action promotes the well-being of the system's intended users or alleviates the suffering of others when used, while the relevant moral reasons against performing some wrong action might have to do with the way in which that particular action is harmful. Arpaly takes acting for the relevant moral reasons to be an agent's desire to perform actions with the appropriate right-making features and to perform actions without wrong-making features. For example, a computer engineer might be praise-worthy for implementing an established ethical design practice when developing automated systems in order to benefit her users and to prevent any possible harm to them. Arpaly calls this Praise-worthiness as Responsiveness to Moral Reasons (PRMR), and she calls the motive from which praise-worthy actions arise good will.

The PRMR specifies two primary epistemic capacities central to the assessment of an entity's moral-worthiness. The entity must have a capacity for reasons responsiveness (i.e., the capacity to act for the substantive features of the action) and the capacity to execute actions motivated by those reasons. Throughout this paper, I have assumed that many machines have the capacity to execute morally significant behaviors, which is relevant to the second primary capacity, but it remains to be seen as to whether such entities could ever be motivated by reasons in the sense specified by Arpaly. A machine's candidacy for moral accountability, as praise- or blame-worthiness, therefore hangs on the question of whether the machine is ever reason responsive. Reasons responsiveness can be divided into two sub-capacities. First, the entity must be able to perceive the relevant moral reasons, and, second, the entity must be capable of being motivated by those reasons. Afterall, if an entity cannot perceive the relevant reasons then the entity cannot respond to them, and if an entity is unable to be motivated by such reasons (e.g., because it lacks a motivational system of any sort) it cannot respond to them. Any entity that lacks either capacity will not be a candidate for moral accountability because that entity could not be reasonably expected to modify its behavior on the basis of morally relevant reasons for acting or not acting.

This brings me to a discussion of Arpaly's account of blame-worthiness, which adds complexities to any discussion about these sub-capacities. Arpaly claims that her criteria for blame-worthiness are similar to her criteria for praise-worthiness. She suggests that people who do the wrong thing from a deficiency of good will or from ill will are morally blame-worthy for their actions. An agent acts from ill will when she acts from the reasons for which the act is wrong (i.e., malevolent motivations). Harmful actions, motivated by implicit or explicit desires to exclude people on the basis of their identities, are motivated by malevolence because excluding people on these grounds is precisely the feature of the action that makes it wrong. More mundanely, an agent also acts from malevolence

[3] I recognize that the inclusion of praise-worthiness as a form of holding accountable seems somewhat unintuitive. For my purposes, holding accountable has to do with attributing some morally significant outcome to an agent, assessing what that action says about that agent's attitude towards others, and taking action, such as praising or blaming the agent, appropriate to the assessment.

when intimidating, belittling, or berating an opponent - all of which are common bullying tactics. As a result, Arpaly thinks ill willed agents are common and that malevolent motivations are not limited to fundamentally evil agents.

An agent suffers from a deficiency of good will when she is insufficiently responsive to the relevant moral reasons against performing a wrong act (e.g., the fact that the action will harm someone). The deficiency of good will is also known as moral indifference or a lack of moral concern, a typical example of which is someone who blindly pursues profit while entirely indifferent to the harmful impact of that activity on others. For example, companies that have recently fired their AI ethics teams in order to expedite the implementation of new AI systems in their products are insufficiently responsive to the relevant moral reasons when acting [51]. These companies are most likely pursuing profit while maintaining plausible deniability in the event anyone is harmed by their products. An agent is indifferent and not malevolent just as long as the means of acquiring profit is not the harmful activity itself (e.g., poisoning the water supply as a byproduct of a manufacturing process versus working as a paid assassin).

Importantly, apparent moral indifference seems to be a symptom of failing to perceive the relevant moral reasons or failing to be motivated by them, which are the two sub-capacities required for reasons responsiveness. The accountability of entities who are morally indifferent is somewhat contentious because it can be caused by local and global capacities failures. An entity that is morally indifferent due to a global failure suffers from a total lack of the relevant capacity and is therefore not able to be held accountable (e.g., an insect). An entity who suffers from a local failure exhibits the relevant capacity more generally but failed to exercise it in a particular instance (e.g., a sociopath). In such cases, the entity may be accountable, though certain conditions apply. The entity's accountability seems to depend on the nature of the failure, an issue that I will not address in this paper.

V. Epistemic Goals for AI Accountability

In the previous section, I noted that a machine's candidacy for moral accountability would most likely hang on the question of whether the machine is ever reason responsive. Arpaly identifies two main conditions on reasons responsiveness. First, the entity must be able to perceive the relevant moral reasons, and, second, the entity must be capable of being motivated by those reasons.

In order to assess whether an AI ever satisfies these conditions, for the purposes of appropriately assigning blame, for example, to an AI, humans require access to information about the entity's epistemic capacities, which may be a function of the epistemic apparatuses with which that entity is equipped. If it entirely lacks the relevant capacities, then blame will never be warranted. If it exhibits the relevant capacities, then it may well be warranted. Whether that is so will depend on which reasons the entity responded to in acting. AI explainability is important to assessing whether an entity satisfies these conditions, and these, by no means exhaustive conditions, specify epistemic goals for AI explainability and impose practical constraints on adequate XAI. I detail these results in the remainder of this section.

Using an Arpaly-like account of moral-worthiness, we can specify at least three epistemic goals for tools that aim at making AI explainable by identifying the kind of information humans need access to in order to know that an AI is accountable for its behavior and to act on that knowledge.

First, we need to know whether the AI has an input system that approximates the capacity of human perceptual systems, in that it can obtain and process information about the world, including morally relevant facts. This may involve human perceptual capacities such as vision and hearing, but it need not. Generative AI can process images and text in order to synthesize outputs, which potentially gives the AI access to the same sort of information as human perceptual tools. To obtain this information about the AI's perceptual capacities, we need to know how that system works and not simply that some approximately similar system exists, in general, to ensure that the system can obtain the specific category of facts, namely moral facts, that we are interested in.

Second, the level of interpretability must be fine-grained enough to identify which precise features of the action the AI used in selecting a behavior to execute. If we do not know what reasons the AI acted for, then we do not know whether it acted for the right sorts of reasons. This is different from information about whether the AI has the capacity to act for such reasons in the first place.

Third, we must know whether the AI is capable of using the morally relevant reasons as reasons for acting. This concern takes two forms. First, we need to know whether the AI has any sort of motivational system, since the morally relevant facts, in themselves, are irrelevant to our assessment of the AI if it cannot use them to settle the question of which behavior to execute. Whether this is possible is partially determined by the second consideration. We need to know whether the AI can distinguish morally relevant reasons at a granularity sufficient for using them to settle the question of which behavior to execute.

Regarding the first consideration, I am not committed to any particular notion of motivation, other than to say that an agent's motivations-in-acting are the motivations that settle the question of whether to act. Arpaly defines motivation in terms of desire, but the question of what precisely motivates action is, as with all things, contentious in philosophy, and I am not sure that we should limit the content of motivational states to any strict notion of desire. Assuming that information about an AI's motivational system is an explainability requirement, we would have to look to our best theories of motivation to spell out which precise details must be epistemically available. Here, I only mean to specify that we would need access to such information to appropriately call an AI explainable.

Regarding the second consideration, it turns out that humans and AI sometimes organize the world differently, even when presented with roughly the same information (e.g., vision versus video). From a practical standpoint, this could make it harder for us to figure out what features of the world an AI is responding to and whether those features are morally relevant. From a more theoretical standpoint, these differences in how we organize the world could mean that we access information at different levels of granularity, which could impact the ability to use some morally relevant information when engaging in moral reasoning.

On the one hand, some AI fail to make the same distinctions as humans. For example, the AI in an automated vehicle might record video of a group of people walking together and interpret the group as a single bounded, jerrymandered object because the AI only tracks information about the speed and trajectory of the objects recorded in its environment. When two or more distinct objects are close to one another, moving in the same direction, and at the same speed, an AI may have no means of distinguishing between them. So, an AI might have the capacity to obtain information about the presence of an individual person, which is morally relevant to driving scenarios, but lack the predisposition to identify individual people as discreet entities unless they are traveling alone. This could mean that an AI might have difficulty in a Trolley Problem scenario if it cannot tell the difference between one person and a group of five people because it infers only the presence of two distinct objects. Automated vehicles have improved at these sorts of tasks over time but multi-object detection and tracking remains one of the most prominent areas of research for automated vehicles. [52] [53]

Interestingly, on the other hand, an AI could observe patterns or features of the world that humans have not previously discovered, or, at least not previously distinguished or categorized. This is likely because AI are able to process considerably more information than humans and at a faster rate. The ability to make these observations is one of the greatest strengths of AI, and various AI tools have made groundbreaking discoveries in fields ranging from the humanities to mathematics to the biomedical sciences. [54] [55] [56] [57] [58] However, we rarely understand what patterns the AI are observing or what principles they are following, and we need this information to understand whether an AI can distinguish morally relevant reasons at a granularity sufficient for using them to settle the question of which behavior to execute.

To address the problem of AI explainability (i.e., understood in more narrow terms of AI interpretability) research into the field of XAI has yielded tools for conducting post hoc interpretations of an AI's decision-making process. Such tools may be necessary for making AI explainable in a way that allows humans to assess whether any AI could ever be held accountable. These three epistemic goals for AI explainability are also practical constraints on the sorts XAI tools that are suitable for the job. For example, these three goals are likely to rule out XAI tools that use less complex AI models (i.e., for interpreting their more complex counterparts) without also taking an ”autointerpretability” approach, like we see with Anthropic and OpenAI. [59] [60] Such tools may provide humans with some insight into the inner workings of more complex AI models, they may be insufficiently sensitive to be of any use in evaluating AI accountability. [4] It seems that such tools might yield some key information about the more complex AI's capacities, as well as information about the AI's motivational system (were it to have one), but that they are unlikely to yield information about which precise features of the prospective behavior motivated it to execute its behavior.

The former set of considerations might provide information about whether the AI is a potential candidate for being held accountable to its actions, but the latter is necessary for identifying whether the AI is indeed accountable. While the former concern matters to questions of AI moral accountability in that it has to do with ensuring that prospective assessors are actually provided with the information about how the AI works so that blame can be assigned, the latter is more central because it has to do with the possibility of gaining any sort of epistemic access to information about which specific strategies the AI used in executing certain kinds of morally relevant behaviors.

VI. Conclusion

This contribution also intervenes on the practical problem of how to move beyond speculative disagreement on questions about whether accountable AI could ever exist. If an AI is explainable in the sense relevant to AI accountability, then we should be able to tell.

AI pose a problem for humanity. They behave very much like humans, increasingly to the point where we are unable to distinguish them from humans. We currently have the ability to generate audio and video that closely approximates the personality, voice, and appearance of living people, and this technology is only likely to improve with time. What we observe from AI seems very much like a human, but the these behaviors are not generated by the kinds of capacities required for holding a person responsible. Existing AI seem to lack any sort of moral status, and, at present, it seems like the designers of these technologies are responsible for their behavior whether good or ill.

However, given the current trajectory, it seems like some AI will eventually gain moral status as AI systems become increasingly complex. In addition to acquiring the protections due to entities with moral status, they will also acquire the responsibilities due to such entities, such as the responsibility to avoid perpetrating harms. However, the trajectory of increasing complexity likely means that the inner workings of these AI will become increasingly inscrutable to us. This means that AI pose a problem of when it comes to moral accountability. Ideally, we would always hold the appropriate individuals accountable to their actions, but there is a responsibility gap at the point where AI begin to gain moral status. Because gaining access to information about the inner workings of AI is difficult, it is difficult to identify whether the AI is indeed accountable. This lack of information will make it difficult to

[4] There is also considerable concern that existing XAI tools are being used to ”fairwash” existing AI models by misleading users about the fairness of a model. [61]

close the gap. So, I have argued that XAI is central to closing the responsibility gap, and I have proposed at least three goals for tools that seek to make AI more explainable.

I am uncertain of whether Arpaly intended to extend her account to non-human entities like AI with moral status, but I believe her account, or one very similar to it, can plausibly be applied to questions about AI explainability and responsibility. In this paper, I applied Nomy Arpaly's account of moral-worthiness to identify the conditions under which an AI with moral status can be held accountable and argued that this account identifies at least three requirements on the success condition on XAI. Any tools that aim to make AI more explainable should plausibly meet these requirements by making the relevant information available to humans interacting with the AI. If we can gain access to this information, then we can make progress on closing the responsibility gap.

References

[1] L. McNulty, "Incident number 505," *AI Incident Database*, 2023. [Online]. Available: https://incidentdatabase.ai/cite/505

[2] K. Roose, "A conversation with bing's chatbot left me deeply unsettled," 2023.

[3] S. McGregor, "Preventing repeated real world ai failures by cataloging incidents: The ai incident database," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15 458–15 463.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[5] N. Arpaly, *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press, 2002.

[6] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and information technology*, vol. 6, pp. 175–183, 2004.

[7] M. P. R. d. Castro, G. C. Antunes, L. M. P. Marcon, L. S. Andrade, S. Rückl, and V. L. Â. Andrade, "Euthanasia and assisted suicide in western countries: a systematic review," *Revista Bioética*, vol. 24, pp. 355–367, 2016.

[8] C. Field and M. Curtice, "Assisted dying: a review of international legislation," *British Journal of Hospital Medicine (2005)*, vol. 70, no. 5, pp. 280–283, 2009.

[9] C. Grosse and A. Grosse, "Assisted suicide: models of legal regulation in selected european countries and the case law of the european court of human rights," *Medicine, Science and the Law*, vol. 55, no. 4, pp. 246–258, 2015.

[10] P. Hieronymi, "The force and fairness of blame," *Philosophical Perspectives*, vol. 18, pp. 115–148, 2004. [Online]. Available: http://www.jstor.org/stable/3840930

[11] S. Nyholm, "Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci," *Science and engineering ethics*, vol. 24, no. 4, pp. 1201–1219, 2018.

[12] P. Königs, "Artificial intelligence and responsibility gaps: what is the problem?" *Ethics and Information Technology*, vol. 24, no. 3, p. 36, 2022.

[13] D. Shepardson, "Google says it bears 'some responsibility' after self-driving car hit bus," *Reuters*. [Online]. Available: https://www.reuters.com/article/us-google-selfdrivingcar/google-says-it-bears-some-responsibility-after-self-driving-car-hit-bus-idUSKCN0W22DG

[14] S. Pednekar, "Incident number 266," *AI Incident Database*, 2022. [Online]. Available: https://incidentdatabase.ai/cite/266

[15] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[16] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović *et al.*, "Ai explainability 360 toolkit," in *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 2021, pp. 376–379.

[17] J. J. Ferreira and M. S. Monteiro, "What are people doing about xai user experience? a survey on ai explainability research and practice," in *Design, User Experience, and Usability. Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 2020, pp. 56–73.

[18] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, ""help me help the ai": Understanding how explainability can support human-ai interaction," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581001

[19] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[20] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust ai," *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[21] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.

[22] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng *et al.*, "Sleeper agents: Training deceptive llms that persist through safety training," *arXiv preprint arXiv:2401.05566*, 2024.

[23] S. Buss, "The value of humanity," *The Journal of Philosophy*, vol. 109, no. 5/6, pp. 341–377, 2012.

[24] I. Robeyns and M. F. Byskov, "The Capability Approach," in *The Stanford Encyclopedia of Philosophy*, Summer 2023 ed., E. N. Zalta and U. Nodelman, Eds. Metaphysics Research Lab, Stanford University, 2023.

[25] D. DeGrazia, "Moral status as a matter of degree?" *The Southern Journal of Philosophy*, vol. 46, no. 2, pp. 181–198, 2008.

[26] O. Horta, "Why the concept of moral status should be abandoned," *Ethical Theory and Moral Practice*, vol. 20, pp. 899–910, 2017.

[27] E. F. Kittay, "Why human difference is critical to a conception of moral standing: An argument for the sufficiency of being human for full moral status," *The Journal of Philosophy of Disability*, 2021.

[28] J. Sebo, "Agency and moral status," *Journal of Moral Philosophy*, vol. 14, no. 1, pp. 1–22, 2017.

[29] S. C. May, "Moral status and the direction of duties," *Ethics*, vol. 123, no. 1, pp. 113–128, 2012.

[30] M. A. Warren, *Moral status: Obligations to persons and other living things*. Clarendon Press, 1997.

[31] M. Cabanac, "Do animals know pleasure," *Mental Health and Well-being in Animals*, pp. 29–46, 2005.

[32] J. Balcombe, "Animal pleasure and its moral significance," *Applied animal behaviour science*, vol. 118, no. 3-4, pp. 208–216, 2009.

[33] P. Bateson, "Assessment of pain in animals," *Animal behaviour*, vol. 42, no. 5, pp. 827–839, 1991.

[34] T. Dougherty and T. Dougherty, *The problem of animal pain*. Springer, 2014.

[35] H.-J. Glock, "Can animals act for reasons?" *Inquiry*, vol. 52, no. 3, pp. 232–254, 2009.

[36] ——, "Agency, intelligence and reasons in animals," *Philosophy*, vol. 94, no. 4, pp. 645–671, 2019.

[37] T. T. Hills, "Animal foraging and the evolution of goal-directed cognition," *Cognitive science*, vol. 30, no. 1, pp. 3–41, 2006.

[38] B. E. Rollin, "Animal pain: What it is and why it matters," *The Journal of ethics*, vol. 15, pp. 425–437, 2011.

[39] L. U. Sneddon, R. W. Elwood, S. A. Adamo, and M. C. Leach, "Defining and assessing animal pain," *Animal behaviour*, vol. 97, pp. 201–212, 2014.

[40] N. Bostrom, "How long before superintelligence?" *International Journal of Futures Studies*, vol. 2, 1998.

[41] S. Conklin, "Do we have procreative obligations to ai superbeneficiaries?" in *International Conference on Computer Ethics*, vol. 1, no. 1, 2023.

[42] C. Shulman and N. Bostrom, "Sharing the world with digital minds," *Rethinking moral status*, pp. 306–326, 2021.

[43] Y. Maruyama, “Moral philosophy of artificial general intelligence: agency and responsibility,” in *Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings 14*. Springer, 2022, pp. 139–150.

[44] L. Floridi, “Ai as agency without intelligence: On chatgpt, large language models, and other generative models,” *Philosophy & Technology*, vol. 36, no. 1, p. 15, 2023.

[45] A. Ladak, “What would qualify an artificial intelligence for moral standing?” *AI and Ethics*, pp. 1–16, 2023.

[46] V. C. Müller, “Is it time for robot rights? moral status in artificial entities,” *Ethics and Information Technology*, vol. 23, no. 4, pp. 579–587, 2021.

[47] D. DeGrazia, “Robots with moral status?” *Perspectives in Biology and Medicine*, vol. 65, no. 1, pp. 73–88, 2022.

[48] D. J. Gunkel and J. J. Wales, “Debate: what is personhood in the age of ai?” *AI & society*, vol. 36, pp. 473–486, 2021.

[49] E. Schwitzgebel, “Ai systems must not confuse users about their sentience or moral status,” *Patterns*, vol. 4, no. 8, 2023.

[50] ——, “The full rights dilemma for ai systems of debatable moral personhood,” *ROBONOMICS: The Journal of the Automated Economy*, vol. 4, pp. 32–32, 2023.

[51] A. Belanger, “Report: Microsoft cut a key ai ethics team,” *Ars Technica*. [Online]. Available: https://arstechnica.com/tech-policy/2023/03/amid-bing-chat-controversy-microsoft-cut-an-ai-ethics-team-report-says/

[52] A. Kampker, M. Sefati, A. S. A. Rachman, K. Kreisköther, and P. Campoy, “Towards multi-object detection and tracking in urban scenario under uncertainties.” in *VEHITS*, 2018, pp. 156–167.

[53] A. Rangesh and M. M. Trivedi, “No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars,” *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 588–599, 2019.

[54] D. Castelvecchi *et al.*, “Deepmind’s ai helps untangle the mathematics of knots,” *Nature*, vol. 600, no. 7888, pp. 202–202, 2021.

[55] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász *et al.*, “Advancing mathematics by guiding human intuition with ai,” *Nature*, vol. 600, no. 7887, pp. 70–74, 2021.

[56] G. Briganti and O. Le Moine, “Artificial intelligence in medicine: today and tomorrow,” *Frontiers in medicine*, vol. 7, p. 27, 2020.

[57] P. A. Keane and E. J. Topol, “With an eye to ai and autonomous diagnosis,” *NPJ Digital Medicine*, vol. 1, no. 1, p. 40, 2018.

[58] A. Tang, R. Tam, A. Cadrin-Chênevert, W. Guest, J. Chong, J. Barfett, L. Chepelev, R. Cairns, J. R. Mitchell, M. D. Cicero *et al.*, “Canadian association of radiologists white paper on artificial intelligence in radiology,” *Canadian Association of Radiologists Journal*, vol. 69, no. 2, pp. 120–135, 2018.

[59] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information fusion*, vol. 58, pp. 82–115, 2020.

[60] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, “An empirical evaluation of ai deep explainable tools,” in *2020 IEEE Globecom Workshops (GC Wkshps*. IEEE, 2020, pp. 1–6.

[61] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, “Can explainable ai explain unfairness? a framework for evaluating explainable ai,” *arXiv preprint arXiv:2106.07483*, 2021.