AI and Ethics
https://doi.org/10.1007/s43681-023-00332-2

ORIGINAL RESEARCH

# Against the opacity, and for a qualitative understanding, of artificially intelligent technologies

**Mahdi Khalili**[1,2]

Received: 9 December 2022 / Accepted: 16 August 2023


**Abstract**
This paper aims, first, to argue against using opaque AI technologies in decision making processes, and second to suggest that we need to possess a qualitative form of understanding about them. It first argues that opaque artificially intelligent technologies are suitable for users who remain indifferent to the understanding of decisions made by means of these technologies. According to virtue ethics, this implies that these technologies are not well-suited for those who care about realizing their moral capacity. The paper then draws on discussions on scientific understanding to suggest that an AI technology becomes understandable to its users when they are provided with a qualitative account of the consequences of using it. As a result, explainable AI methods can render an AI technology understandable to its users by presenting the qualitative implications of employing the technology for their lives.

**Keywords** Understanding · Opacity · Normativity of Technology · Decision Making · Virtue Ethics · Artificial Intelligence · Explainable AI

## 1 Introduction

In recent years, philosophers of science and ethicists of technology have paid special attention to artificial intelligence (AI). One of their concerns is that AI systems that depend on machine learning techniques instantiated in neural networks produce outputs such that even their makers do not know why a given pattern has been extracted from a given dataset. These AI systems are called "opaque", "non-transparent", "black boxes", or "unintelligible" (see [1, 2], and for a distinction between three kinds of opacity, see [3]). While I particularly consider the opacity of machine learning, what I mean by "opacity" in this paper is a more generic term for *any artificially intelligent technology* that is not understandable *to its users* when the understanding of that technology is morally *relevant*. This opacity raises epistemological and ethical questions. The epistemological question concerns the question of whether and how we can understand AI technologies. The ethical question is about the normative implications of the opacity/understandability of AI technologies for our lives.

In a recent article, Claus Beisbart and Tim Räz [4] suggest that philosophers of science should engage with these tasks concerning the interpretability of artificial intelligence: they should "(i) clarify the notion of interpretability, (ii) explain the value of interpretability, (iii) provide frameworks to think about interpretability, and (iv) explore important features of it to adjust our expectations about it." Beisbart and Räz bracket moral concerns in this suggestion of theirs. I think, however, that philosophy of science and technology ethics cannot be separated in properly addressing the issue of understanding in AI, and thus philosophers of science (and epistemologists) should develop relevant accounts of understanding that consider moral concerns from the beginning, as I shall do in this paper. Section 2 will connect the epistemic understanding that we need to have about AI technologies with the role of understanding in having "moral perspective". Furthermore, Sect. 3 will suggest that we need a qualitative form of epistemic understanding, because it can provide the consequences of using AI technologies for our moral lives. Thus, this paper integrates the epistemological and ethical aspects of the problem of understanding about

✉ Mahdi Khalili
mahdi.khalili@uni-graz.at; mahdi.khalili@ipm.ir

[1] Department of Philosophy, University of Graz, 8010 Graz, Austria

[2] School of Philosophy, Institute for Research in Fundamental Sciences, Tehran, Iran

Published online: 04 October 2023

AI technologies (on the connection between ethics and epistemology of AI, see also [5, 6]).

The concepts of transparency, interpretability, explainability, and understandability are all used in the philosophical and scientific literature on machine learning. Among them, however, the notion of understandability is more fundamental. "Transparency" has a somewhat metaphorical meaning: a thing such as a box that allows you to see its inside is "transparent". But if you can see the inside of the box and do not understand what is inside, it is not really transparent. Accordingly, for a thing to be transparent the thing should be, in some way and to some extent, understandable. Also, societies place a high value on "transparency" because of the profound significance of the human faculty of understanding. People want to understand social and political courses of action that may influence their lives, so they demand transparent procedures. "Interpretation" is also nothing but an extension of understanding. One can explicate an implicit or a succinct understanding of something in more detail so as to offer a relevant interpretation. Finally, "explanatory" strategies and techniques that render AI "explainable" are desirable insofar as these tools can facilitate the aim of understanding (see [7]). Because of its importance, the notion of understanding is central to this paper.[1] I use "opacity" as the opposite of understanding. Technologies that are not (completely) understandable are (to some extent) opaque.

The structure of the paper is as follows. Section 2 presents my main argument against opaque AI technologies, according to which they are suitable for those who do not care about realizing their moral capacity. This section also clarifies that my argument is not directed against opaque AI technologies per se, but against using them in decision making processes that are relevant to our moral lives. Section 3 suggests that the kind of understanding that we need to have about technologies is qualitative, and accordingly XAI methods, which aim to make AI systems understandable to humans, should provide a qualitative account of how using a technology influences our lives. As a result, even if AI remains opaque quantitatively/computationally, by using XAI methods that provide qualitative understanding, we may still reach some level of understanding needed for our ethical purposes.

[1] Despite Carl Hempel's ([8], p. 413) early dismissal of 'understanding', which he considered to be an insignificant by-product of explanation, understanding is nowadays a central concept in philosophy of science. Moreover, understanding is an essential capacity of *human* characters, and as Sect. 2 will clarify, one needs it to be virtuous. AI systems are problematic when they are designed without taking our human ability to understand into consideration.

## 2 Against the opacity of AI technologies

My concern about opaque technologies, in particular, opaque AI, can be set out in the form of an argument as follows.

(1) Opaque AI technologies are suitable for users who do not care about the understanding of decisions (made by means of these technologies) that influence their lives.
(2) Those who do not care about the understanding of these decisions do not realize their moral capacity.

Therefore, opaque AI technologies are suitable for users who do not realize their moral capacity. This also implies that these technologies are inappropriate for those who prioritize realizing their moral capacity.

Although my argument against opacity may be more generally applicable to using any opaque technology, I particularly intend to apply it to artificially intelligent technologies inasmuch as we outsource our *decisions* to them. Thanks to our capability of practical reasoning, we tend to critically analyze the reasons behind decisions that are relevant to our lives. However, the problem comes up when the reasons behind these decisions are unclear, and so critical assessments of them cannot be made. The following two subsections explain the premises of my argument in detail.

### 2.1 The first premise

The first premise assumes the *normativity* of technology. That is, for a functional system to work in a stable and reproducible manner, certain social and technical contexts *should* be established. I follow Hans Radder ([9], chapter 2) on the definition of technology and its inherent normativity. He characterizes "a (type of) technology as a *(type of) artifactual, functional system with a certain degree of stability and reproducibility*" ([9], p. 47). A collection of mutually interacting material entities constitutes a system. An artifactual, functional system is produced by humans to serve their purposes. A technological system is stable in the sense that it carries out its function across different situations and times, and is reproducible in the sense that it belongs to a type of systems, all of which can exhibit the same function. The normativity of technology can be understood as a result of this definition. For a type of technology to function stably and reproducibly, the relevant techno-social context *should* be appropriate. Given the necessity of this "should", technology is thus inherently normative: "technologies are inherently normative because their stable and reproducible realization in some region of space and time requires that the people in that region should behave in such a way as to enable, and not disturb, the intended functioning of the technology" ([9], p. 58). The claim that technologies make certain requirements on people should be comprehended at

the same time at the personal and social level. The stable and reproducible realization of a technology requires us to have certain personal characteristics and certain social-political conditions to be established. AI technologies are obviously kinds of technology, so they are normative as well, and our personal characters and socio-technical conditions must be appropriately changed to fit them if those technologies are to realize their full potential.

Now my claim is that people who do not care about the understanding of the reasons for decisions that are made by opaque AI technologies are among the conditions for the full realization of these technologies. Humans typically ask for the reasons for decisions that impact their lives. However, the questions raised by humans interested in understanding opaque AI technologies cannot be answered, and if the potential/intended users insist on their questions before using these technologies, they will indeed disturb the proper functioning of these technologies. On the other hand, those who do not care to understand reasons for decisions are among the *suitable* conditions for these opaque technologies to function effectively. In this sense, in order to use opaque AI technologies their users are discouraged from understanding the decisions made by means of these technologies, and thus they are discouraged from developing their capability of understanding. The users can enjoy the efficiency of these opaque technologies without going to the trouble of asking serious questions about the reasons behind the decisions made by means of them.

The first premise can also be supported by an idea of Michel Foucault expressed in his *Discipline and Punish* that power and the material artefacts that exert power (for example, the Panopticon prison) constitute certain subjects. For Foucault, ideas, artefacts, and institutions provide the world in which human subjects live, and through them, power shapes human subjectivity. "Technology can be seen as one of these sources of power that help to shape the subject" ([10], chapter 4). An implication of this Foucauldian analysis of subject-constitution for our discussion is that opaque AI technologies constitute subjects who do not concern themselves about the reasons for decisions. In particular, the efficiency of artificially intelligent technologies establishes (seemingly legitimate) powers of shaping subjects/characters that willingly devolve their reasoning abilities to machine algorithms and techniques.

According to the first premise, opaque AI technologies are suitable for users who do not care about understanding reasons. These technologies make it less likely or more difficult for the users to understand the reasons behind the decisions made by these technologies. This claim of mine does not imply that we always have to understand the complicated systems behind technologies. We barely understand how an LED screen works or many of us use cars without understanding how they work, but we can use these technologies for morally acceptable purposes. Similarly, one can use an AI technology without understanding the complicated, computational system behind it. Subsection 2.3 will clarify that my argument is not against opaque AI technologies themselves, but rather against using them in decision making contexts that are relevant to our moral lives. Section 3 will also clarify that the kind of understanding that is essential for my argument in this section is qualitative. We need to evaluate the qualitative implications of using artificially intelligent technologies for our lives.

## 2.2 The second premise

The second premise explains why the constitution of agents who do not care about the understanding of decisions that are relevant to their lives is undesirable. Such agents do not realize their capacity of practical wisdom, or what Shannon Vallor ([11], chapter 6) calls "technomoral wisdom", in which "technomoral virtues"—i.e., virtues that are necessary to live a good life in the age of emerging technologies—are integrated. One of these technomoral virtues is "moral perspective", which Vallor defines "as a reliable disposition to *attend to, discern, and understand moral phenomena as meaningful parts of a moral whole*" ([11], p. 149). A person who is insensitive to the understanding of reasons for decisions does not discern or understand moral phenomena appropriately. This person can neither pay serious attention to morally relevant factors, nor grasp the importance of these factors in the broader *context* of a decision. Indeed, the moral perspective explains the connection of "understanding" and "practical wisdom" by highlighting the key role of the former in the latter. Moreover, because the moral perspective is "an essential disposition of a virtuous person" (2016, pp. 149–150), those who have an insufficient moral perspective are unable to practice other virtues such as justice, honesty, care, and civility. As a result, those who lack adequate understanding cannot cultivate (technomoral) virtues.

As a result of this virtue framework, the critical understanding of decisions is central to the moral lives of humans. Thus, opaque artificially intelligent technologies are undesirable insofar as they undermine practical/technomoral wisdom. In a recent talk entitled "Thinking outside the black box: AI and the shrinking space of moral reasons", Vallor ([12]) reaches a similar conclusion. She says: "the personal and public space of moral reasons is contracting as the power and socio-economic utility of sophisticated machine algorithms expands". This conclusion is drawn on the basis of her use of Wilfrid Sellars's and Robert Brandom's concept of the "space of reason". My argument does agree with this use, but I have a further point to make. The first premise of my argument sheds light on a deeper implication of the shrinking space of moral reasons: the contraction of the

moral scope of the characters deploying these technologies. It is not the case that highly moral characters can live, albeit with difficulty, in a narrow moral space of reason. Rather, characters themselves become less morally sensitive in such a narrower space.

The logical result of both premises is that opaque AI technologies are suitable for users who do not realize their moral capacity. This result simply implies that these technologies are inappropriate for those who prioritize realizing their moral capacity. As a consequence, the demand for understandability of AI technologies is necessary for people and communities to take care of their practical wisdom. In order to have a society with practically wise personalities, we should avoid using (and designing) opaque technologies in our decision making processes and should take steps to use (and design) understandable ones.

## 2.3 Decision making and use contexts

This subsection clarifies that my argument is not against opaque AI technologies themselves, but rather against using them where our understanding of these technologies is morally relevant and important to the implications of decisions made by means of them for our lives. In this regard, two terms are key: decision making and use contexts.

### 2.3.1 Decision making

The main difference between AI technologies and other examples of technologies, such as LED screens, is that the former are the workhorses of decision making. For this reason, as I mentioned earlier, my argument particularly applies to AI technologies, which are designed to systematically extract information (from data) that provides correlations, predictions, or interpretations that support decision making processes. We need to understand decisions that influence our lives, in particular decisions that are laden with moral concerns. In many cases of AI technologies, such as facial recognition (as a technology of surveillance), stock market predictions, social media analyses, medical methods of diagnosis, signature and handwriting analysis, university admissions and job placements, understanding reasons behind decisions made by AI is completely relevant to our lives.

Consider job placement processes, in which some level of explanation seems to be necessary all the time. For a more concrete example, consider a recent piece of news, according to which Iran's Ministry of Science, Research, and Technology will deploy an AI technology in its faculty recruitment process. According to the Deputy Head of the Planning and Executive Center of the Ministry, "with the activation of the AI system, better monitoring will be done in recruiting faculty members, which prevents making arbitrary judgements" ([13], my translation). It remains unclear how this system may prevent arbitrary decisions and actions, in particular those amounting to discrimination, while bias might be even reinforced by AI (see [14], chapter 9). Apart from this, this new system is obviously relevant to the lives of many applicants, and thus it should always be understandable. The applicants need to know the criteria behind the decisions made by the system and to assess whether the decision making process is fair.

My emphasis on decision making processes coheres with my suggestion in the next section that an AI technology can become understandable if a *qualitative* account of the consequences of its use can be provided. We should enjoy qualitative understanding to make decisions that carry good implications for our lives. I will come back to the "qualitative" aspect of the job placement example in Sect. 4.

### 2.3.2 Use contexts

To highlight the importance of use contexts in the ethical assessment of technologies, I would like to compare the conclusion of my argument with a similar claim made by Nathan Colaner [15]. He argues that unexplainable systems are dehumanizing, because they threaten (I) our participation in decision making processes, (II) our knowledge of how AI systems influence us, and (III) our opportunity to actualize ourselves through making decisions. These claims are compatible with, yet different from, mine. In particular, Colaner's third claim that "people are not able to fully actualize themselves unless they are able to meaningfully participate in the decision making procedure" (2021, p. 6) is comparable to the conclusion of my argument. Still, there is a main difference, which I shall clarify in what follows.

He claims that explanations can be "intrinsically" valuable. That is, apart from their instrumental usefulness in promoting other values—such as fairness, trust, accountability, and manageability—they can be valuable in themselves: "By definition, instrumental approaches seek an explanation to attain some other value. It is also possible to argue that explanations are valuable in themselves—intrinsically so—rather than only deriving their value from the hope that they will help us gain some other value" ([15], p. 3). I think, on the other hand, it is highly problematic to assume that an explanation can be valuable in itself, regardless of its *use context*. An explanation that is used for the purpose of a terrorist attack cannot be good merely because the explanation increases the terrorists' participation in decision making processes, or increases their relevant knowledge or their opportunity of actualizing themselves. It is always necessary to consider how this explanation is used in the whole context. That is to say, the explanation can only be good when its implications in a context of use are ethically acceptable.

It is, thus, untenable to consider understandable AI to be "intrinsically" valuable. In general, the value of a (type of)

technology depends on its *use* as an instrument in human *contexts*. The functional role of a technological artefact, including an AI technology, implies that the artefact is valuable if its function in a context of use supports a moral value (cf. [16]). In this sense, technological artefacts can only be *instrumentally* valuable. Accordingly, the (undesirable) effect of an opaque AI technology on the human condition is indeed produced as a result of its use as an instrument in context, and therefore it does not hold water to claim that opaque AI is intrinsically bad "regardless of whether the outcome of having explainable systems is desirable" ([15], p. 2). My argument entails that the characters who do not concern themselves with the understanding of reasons for decisions are indeed the results of employing opaque AI technologies. Thus, a difference between the conclusion of my argument and Colaner's third claim concerns what I stated earlier: I do not argue against opaque AI technologies themselves, but against using them where morally relevant factors rely on our understanding of these technologies.

## 3 For a qualitative understanding of AI technologies

The previous section has argued that opaque AI technologies are suitable for users who do not realize their moral capacity. The kind of understanding that is important for the sake of my argument is qualitative because, in order for the users of AI technologies to decide and act wisely, they need to evaluate the qualitative consequences of using these technologies for their lives. On the other hand, to make morally and practically wise decisions, they do not necessarily need to understand the complicated, computational structure of that technology.

The present section clarifies this qualitative account of understanding and suggests that XAI methods can make an artificially intelligent technology understandable inasmuch as they provide qualitative consequences of using that technology. Accordingly, these methods should render the decision making process of an artificially intelligent technology qualitatively understandable to its users.[2] Before presenting my preferred view of understanding, I evaluate a related claim made by John Zerilli et al. [18] in the next subsection. Although I agree with some of their concerns, my view can overall be considered as an alternative to theirs.

### 3.1 Limitation of intentional stance explanation

Zerilli et al. [18] argue that the justification of action at the level of practical reason is preferable to that of action at the level of "the architectural innards" of a decision tool. According to them, we justify human action based on folk concepts, attitudes, beliefs, and desires rather than on the physical or biological structures of the brain. In like manner, a justification of neural networks that implement deep learning algorithms should be provided at the level that Daniel Dennett [19] calls the "intentional stance": "This is the stance from which we understand ordinary human behaviour and engage in practical reasoning" ([18], p. 669). It is reasonable that our understanding of the physical structure of the brain and the architectural innards of neural networks are usually irrelevant to the justification of behaviors of humans and machines. However, it remains unclear why we should assume that different explanation styles—such as input influence-based explanations, demographic-based explanations, case-based explanations, and sensitivity-based explanations [20]—employ intentions, attitudes, beliefs, and desires in their outputs. For instance, consider these two sensitivity-based explanations.

> If 10% or less of your driving took place at night, you would have qualified for the cheapest tier.
> If your average miles per month were 700 or less, you would have qualified for the cheapest tier.
([20], p. 6)

In what sense are the concepts used in these two sentences expressed at the intentional level? According to the Dennettian view of Zerilli et al., the answer would be thus: *as if* the artificially intelligent machine "believes" that these two sentences are the case. In this answer, the intentional stance is attributed to machines metaphorically. But I do not think that this metaphorical attribution is necessary. In what follows I suggest an alternative view, according to which the decisions made by means of machines should be *qualitatively* understandable. The concept of qualitative understanding conveys this intuition that our understanding of AI systems need not depend on computational descriptions of their architectural innards. Thus, my alternative view agrees with Zerilli et al.'s account that the justification of action at the level of the architectural innards of a decision tool is neither necessary nor helpful. However, my view does not need to assume explanation styles to (metaphorically) have intentions, attitudes, beliefs, and desires.

Qualitative understanding can usually be stated in simple terms, and in this regard I also share Zerilli et al.'s concern that excessively detailed and lengthy explanations are not usually helpful in providing understanding. Despite their account, at the same time, qualitative understanding should

[2] In addition to being a solution to the problem of opacity, XAI can improve the exploratory potential of machine learning and data-driven scientific inquiry: "Explainable AI is a promising new tool for scientific exploration, and is likely to profoundly impact data-driven scientific research" ([17], p. 237).

not necessarily be stated in terms of folk concepts. It may also be stated by using scientific and technical terms. Understanding is agent-dependent in the sense that a system is understandable or opaque *to human agents*. In other words, understanding is obtained by different stakeholders, who may expect different forms of understanding (see [21]). In particular, developers may need technical explanations to gain relevant understanding of a given system. In the argument of Sect. 2, I paid special attention to the (end) users of technologies, who typically prefer non-technical explanations. However, this does not mean that qualitative explanations cannot use technical terms more generally, and thus it would be unreasonable to restrict an explanatory understanding to an untechnical one expressed just by folk concepts. The next subsection elaborates the general elements of the qualitative kind of understanding I advocate. To do this, I draw on discussions on scientific understanding.

### 3.2 Drawing on scientific understanding

The literature on scientific understanding concerns the understandability of *theories to scientists*, and so it is different from our discussion on the understandability of AI *technologies to their users*. Still, it is important to note that, generally speaking, scientific practices are not completely different from other human activities. They tend to be more systematic (see [22]), but not to be "entirely distinct or disconnected from the range of abilities that support ordinary human life" ([23], p. 18). As a result, scientific understanding is not incomparable with other forms of understanding incorporated in practices such as solving everyday problems, negotiation, management, conflict resolution, and judicial decision making. This view paves the way for the use of "scientific" understanding as a guide in other contexts (and vice versa: for the use of "non-scientific" understanding as a guide in scientific contexts). In particular, what this subsection suggests is that the project of developing XAI methods could *draw inspiration from* how scientists make opaque phenomena or models understandable.

There are several theories/models in science that are predictively successful, although they are faced with the "black box" problem. Scientists usually formulate models to make these black boxes understandable. For example, computer simulations make weather predictions based on complex calculations that are hard to understand, but to make them understandable, meteorologists have developed "PV thinking", whose goal "is to provide qualitative understanding … by means of a relatively simple picture" ([24], p. 106). Another example is the construction of so-called "bag models" to make quantum chromodynamics (QCD) understandable (2017, pp. 112–113). I see the project of developing and using XAI methods as being in a *similar* vein. XAI developers desire to make opaque AI understandable by constructing simple understandable models.

In spite of this similarity, we should be cautious about applying accounts of scientific understanding to AI, to which all aspects of these accounts are not applicable. In particular, I do not claim that ordinary people should always have a theoretical, model-based sort of understanding. In what follows, among philosophical accounts of scientific understanding (see [25]), I specifically refer to Henk de Regt [24]. He asserts that his theory applies to the natural sciences, and that "further research should reveal to what extent the theory possesses a wider validity" ([24], p. 11). I do not claim that his theory is applicable in detail to artificially intelligent technologies. I think, instead, that there are some elements in his account that are relevant to our discussion in this paper.

In the first place, the point I made in the introduction—that understanding is a *human* capacity—is taken seriously in his account: "One can use the term 'understanding' only with—implicit or explicit—reference to human agents" ([24], p. 19). According to him, "understanding, in contrast to explanation, necessarily involves a subject. Thus, if information is considered as contributing to understanding, it must be in principle accessible to the understanding subjects; persons who use the explanation must be able to know or grasp the information" ([24], p. 84). This does not imply that understanding should be reduced to a merely subjective feeling, or to "eureka" or "aha" experiences, which is neither a necessary nor a sufficient condition for understanding ([24], pp. 20ff.). But it simply means that the understanding of phenomena, models, and systems should be obtained *by*, and be *within*, the human capacity of understanding.

In the second place, De Regt's theory of understanding highlights the role of *contexts* (see chapter 4 [24]). He convincingly argues that there are no universal standards of understandability. That is, standards such as causality, visualizability, and unifying power apply only to certain cases. Thus, "there is a variety, even a plurality, of explanatory strategies to attain the aim of understanding" ([24], p. 85). Similarly, I suggest that there may be various methods to render AI technologies understandable. One could understand a technology based on the causal relations between the different features of its system, or based on a visual diagram that explains its operation, or based on the resemblance of its workings to a more general, unified picture/theory, to name but a few methods. This pluralist account of understanding is more useful than a monist account that confines our capacity of understanding to within a certain line of reasoning.

The contextual theory of understanding is also in line with my ideas in the previous sections. The relevance of moral factors, and their importance, rely on the context of a decision. For this reason, "moral perspective" is always contextual, in the sense that we should discern particular aspects of the context to have a morally adequate perspective.

Furthermore, as said in my criticism of Colaner, an XAI method is morally valuable if its function in a context of use supports a moral virtue. Therefore, moral perspectives, and moral virtues that are relevant in a context, may specify what kind of understanding we need, and in turn what kind of standards of understanding can be employed.

The third, and most important, feature of De Regt's theory of understanding concerns its emphasis on the *qualitative* property of understanding. His Criterion for the Intelligibility of Theories follows: "A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations" ([24], p. 102).

Two steps should be taken to make this criterion useful for the understandability of AI technologies. The first step is to clarify how it can help us to establish a criterion for the understandability of *artificially intelligent technologies.* The second step is to specify which aspect(s) of this criterion is important *for* agents who are not necessarily scientists. We should be careful to apply only the relevant aspect(s) of this theory of understanding to AI technologies.

Regarding the first step, it should be clarified what it means to possess a qualitative recognition of an AI technology without having exact calculations. My response is that an AI technology can be recognized qualitatively when our recognition of it does not depend on the *computational* processes that take place at the level of its architectural innards (so, in this regard, I am sympathetic to Zerilli et al.). Causal reasoning, visual representations of significant mechanisms, and discovering continuity/resemblance between the AI technology and other understandable systems can provide kinds of qualitative understanding, but there may be several other conceptual tools. Understandability is a pragmatic and context-dependent property (see [24], p. 141), so the achievement of the understandability is related to the characteristics of that technology, its contexts of use, and the questions and purposes of those to whom the technology should be intelligible.

Regarding the second step, I would think that the core of De Regt's criterion, that is, the proposal that understanding is "qualitative", is the aspect that is particularly relevant to the understandability of AI technologies. According to his criterion, agents "can recognize qualitatively characteristic consequences of T without performing exact calculations", but what might be the nature of the qualitative *consequences* of an AI technology? This question may have different meaning for, and thus its answer will depend on, the agent/stakeholder to whom the technology should be understandable. For instance, AI scientists and developers should possess some qualitative sense of how the system produces its outputs. The (end) users, on whom I concentrate in this paper, should understand how the technology will affect their courses of actions and plans. In this regard, and as far as the argument of Sect. 2 is concerned, I would like to highlight the point that the consequences are not merely epistemological, but can be moral as well. Virtuous characters possess prudential judgment, that is "*the cultivated ability to deliberate and choose well, in particular situations, among the most appropriate and effective means available for achieving a noble or good end*" ([11], p. 105). Although virtue ethics is in tension with merely consequentialist normative ethics, having the ability in prudential judgment requires being able to consider some foreseeable consequences of a decision. Prudent users examine the moral consequences of using an AI technology in order to see whether it is an appropriate tool to achieve good purposes. As a result, these users need to be aware of the morally relevant, qualitatively characteristic consequences of using the technology. XAI designers and developers should accordingly provide the users with this qualitative kind of understanding. They may themselves need explanations that deploy technical terms. However, first, these explanations may be achieved even without having the detailed knowledge of the complex computational systems of opaque technologies. And second, as far as my argument in this paper is concerned, the purpose of their explanations is to ultimately shed light on the consequences of using the AI technology for the users' lives.

## 4 Conclusion

Let us consider the job recruitment example once again. Further to the argument of Sect. 2, in the use context of an opaque job recruitment technology, the users accept the outputs of the system without raising questions, challenges, or complaints that disturb the functioning of that technology. Thus, several morally relevant issues remain unaddressed unless the new technology becomes understandable. The questions of whether this technology is fair, which characteristics of applicants are more important for the decision process of this new system, whether these characteristics are weighted in a way appropriate to the job requirements, and similar questions, cannot be dealt without having a qualitative understanding of the technology. On the other hand, XAI methods render such a technology understandable insofar as they can provide the users with the awareness of the qualitative consequences of its use. Applicants want to be aware of, for instance, what parts of their CVs may increase or decrease their chances of acceptance, and whether the new technology justly selects the most qualified candidates. They do not need to know the complex computations it performs. They *understand* the technology when they can explore the concrete impact of its implementation on how their personal and professional characteristics are assessed.

For another example, consider Hang the DJ, the fourth episode of the fourth season of the *Black Mirror* series. In this episode, Amy and Frank use an artificially intelligent dating technology that pairs up couples but puts an expiration date on relationships in order to finally, through numerous relationships, find their ultimate compatible partners. But when they ask about the decisions made by this technology, its answer is merely: "everything happens for a reason". This response indicates that the technology is opaque,[3] which is frustrating for the users and weakens their trust in the technology. My argument in this paper is primarily concerned neither with this frustration nor with the trust of the users. I instead suggest that Amy, Frank, and other users of this dating technology do not enjoy the opportunity of realizing their capacity of understanding. This opaque technology is just suitable for those who do not care about the realization of this capacity. Furthermore, the kind of understanding that a user of such dating apps need is qualitative. The user expects to understand why s/he is paired up with P (the suggested partner), and why a specific expiration date is calculated. If the user seeks a long-term relationship, for instance, such an explanation is required: "signs of compatibility cannot be found between you and P, because you do not have common goals and interests in terms of where to live, how to spend spare times, and how many children to have. In similar cases, the disagreements between partners have not been settled amicably. The prediction of the system, based on the analysis of other relationships, is that your relationship will not last very long." Or: "you and P match, because your physical characteristics, approaches in solving conflicts, sexual expectations, and plans for work and life fit together. In similar cases, the relationship has happily lasted for many years. For this reason, you should stay in this relationship." Or this less certain piece of advice: "According to the analysis of the system, you should think seriously about the following three questions (i) do you have a realistic plan to grow together, (ii) do you really allow your partner freedom to do what you don't like, and (iii) do you respect your partner's family, even if you do not enjoy spending time with them? Your relationship will be fragile unless you can address the challenges relevant to these three questions." These explanations help the user to understand why a piece of advice is offered and how (the reasons behind) the advice can influence their decisions and actions. They can then accept, reject, or just partly follow the advice, and so through an active engagement with the technology, they can still realize their capacity of understanding. These simple examples serve to illustrate how a qualitative form of understanding looks like. In practice, however, more complex situations require more sophisticated and detailed, yet still qualitative, explanations.

This paper has argued that human moral capacity cannot easily be realized in the context of using opaque artificially intelligent technologies. It has then suggested that when an opaque AI technology is employed in decision making processes, it must be augmented with pertinent XAI methods that provide qualitative understanding. XAI methods encompass different approaches, some of which are explicitly designed to explain opaque AI models including machine learning. These methods can achieve the goal of enhancing the understandability of an opaque AI technology if they enable its users to grasp the qualitative implications of implementing the technology in their daily lives.

**Acknowledgements** I would like to thank Federica Russo, Saeedeh Babaii, and the two reviewers for this journal for their constructive comments. I also extend my gratitude to the organizers and participants of the following conferences, in which I presented earlier versions of this paper: Explanatory AI: Between Ethics and Epistemology (TU Delft, May 2022), The Philosophy of Data Science (Frankfurt School of Finance and Management, June 2022), Neurotechnology Meets Artificial Intelligence (LMU Munich, July 2022), and Social Justice and Technological Futures (Tübingen, May 2023).

**Author contributions** This is a single-authored paper.

**Funding** Open access funding provided by University of Graz.

**Availability of data and material** Not applicable.

## Declarations

**Conflict of interest** Not applicable.

**Consent for publication** Not applicable.



## References

1. Burrell, J.: How the machine thinks: understanding opacity in machine learning systems. Big Data Soc. **3**(1), 1–12 (2016)
2. Müller, V. C.: Ethics of artificial intelligence and robotics. In *The Stanford encyclopedia of philosophy*, edited by Edward N. Zalta.

[3] The setting of the movie is unusual, and eventually at the end of the episode, the viewer understands that Amy and Frank have been in a simulated world. Here, I do not discuss the 'simulation' aspect of the characters, but just the opacity of the technology they use.

https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/ (2021a).
3. Müller, V. C.: Deep opacity undermines data protection and explainable artificial intelligence. In *AISB 2021 Symposium Proceedings: Overcoming Opacity in Machine Learning,* 18–21 (2021b)
4. Beisbart, C., Räz, T.: Philosophy of science at sea: clarifying the interpretability of machine learning. Philosophy Compass (2022). https://doi.org/10.1111/phc3.12830
5. Durán, J.M., Sand, M., Jongsma, K.: The ethics and epistemology of explanatory AI in medicine and healthcare. Ethics Inf. Technol. **24**, 42 (2022). https://doi.org/10.1007/s10676-022-09666-7
6. Russo, F., Schliesser, E., Wagemans, J.: Connecting ethics and epistemology of AI. AI & Soc. (2023). https://doi.org/10.1007/s00146-022-01617-6
7. Páez, A.: The pragmatic turn in explainable artificial intelligence (XAI). Mind. Mach. **29**(3), 441–459 (2019)
8. Hempel, C.G.: Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press, New York (1965)
9. Radder, H.: From commodification to the common good: Reconstructing science, technology, and society. University of Pittsburgh Press, Pittsburgh (2019)
10. Verbeek, P.-P.: Moralizing technology: Understanding and designing the morality of things. University of Chicago Press, Chicago (2011)
11. Vallor, S.: Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, Oxford (2016)
12. Vallor, S.: Thinking outside the black box: AI and the shrinking space of moral reasons. [Video]. https://www.youtube.com/watch?v=WzZv8mvZGPM (2022, February)
13. Anonymous: The introduction of AI in the process of recruiting academic staff of universities in order to eliminate arbitrary recruitment. Irna (2022 April 9). https://irna.ir/xjJ2Hc
14. Coeckelbergh, M.: AI Ethics. MIT Press, Cambridge, MA (2020)
15. Colaner, N.: Is explainable artificial intelligence intrinsically valuable? *AI & Society*: 1–8 (2021)
16. Van de Poel, I., Kroes, P.: Can technology embody values?. In: *The moral status of technical artefacts*, pp. 103–124. Dordrecht, Springer (2013)
17. Zednik, C., Boelsen, H.: Scientific exploration and explainable artificial intelligence. Mind. Mach. **32**(1), 219–239 (2022)
18. Zerilli, J., Knott, A., Maclaurin, J., Gavaghan, C.: Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol **32**(4), 661–683 (2019)
19. Dennett, D.: The intentional stance. MIT Press, Cambridge, MA (1987)
20. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. pp. 1–14 (2018). https://doi.org/10.1145/3173574.3173951
21. Zednik, C.: Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol **34**, 265–288 (2021)
22. Hoyningen-Huene, P.: Systematicity: The Nature of Science. Oxford University Press, New York (2013)
23. Chang, H.: *Realism for Realistic People. A New Pragmatist Philosophy of Science*. Cambridge: Cambridge University Press (2022)
24. De Regt, H: Understanding scientific understanding. Oxford University Press, Oxford (2017)
25. Grimm, S. R., Baumberger, C., Ammon, S.: *Explaining understanding: New perspectives from epistemology and philosophy of science*. Routledge (2017)