Organon F 32 (4) 2025: 380–401 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2025.32401

RESEARCH ARTICLE

The Dual Nature of Explicability in AI Ethics

Hyundeuk Cheon*

Received: 2 October 2024 / Revised: 9 April 2025 / Accepted: 13 April 2025

Abstract: Despite the significance of explicability in AI ethics, the principle of explicability remains subject to several unresolved issues, including its moral status, purpose, and the recipients of explanations. First, this paper proposes treating explicability as a prima facie duty to make machine learning algorithms explicable. Second, the dual nature of explicability is highlighted. It is claimed that explicability is for the warranted trust of decision-recipients in the algorithmic decisions as well as for enhancing the autonomy of decision-makers.

Keywords: Algorithm; explicability; explainability; trust; trustworthiness; autonomy.

1. A Call for the Principle of Explicability

Machine learning algorithms (hereafter, ML algorithms) are increasingly being utilized in complex, real-world decision-making, which is often of ethical significance. While the obvious cases are autonomous weapon systems and self-driving cars, these algorithms are also used in ordinary decision-making, from reviewing job applications, assessing loan applications, and approving parole to predictive policing. They are designed to help allocate

* Seoul National University
https://orcid.org/0000-0002-7569-7776
✎ Seoul National University, Department of Science Studies, Building 25-419, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826
✉ hdcheon@snu.ac.kr

social services, decide on promotions or terminations, determine credit scores, or estimate a person's risk of committing crimes. With rapid technological development, there are growing concerns about the ethical and responsible uses of AI algorithms. In particular, the demand for explainability or transparency in AI ethics has attracted considerable attention. For example, European Union's *General Data Protection Regulation* is often interpreted as incorporating the 'right to explanation' when someone is affected by automated decision-making (Goodman and Flaxman, 2017). The Future of Life Institute (2017) declared the Asilomar AI principles, which emphasized the transparency of algorithms when it causes harm or is involved in judicial decision-making.[1] Microsoft's CEO Satya Nadella (2016) also calls for a similar requirement of intelligibility in terms of understanding "how the technology works and what its rules are."[2]

Luciano Floridi, one of the leading figures in AI ethics, suggested that similar principles - the principles of explainability, transparency, interpretability, and accountability - can be incorporated under an overarching principle, what he called the principle of explicability (Floridi et al., 2018; Floridi and Cowls, 2019). Floridi and his colleagues put forth the notion of explicability as encompassing both the epistemological sense of intelligibility and the ethical sense of accountability. The former concerns an answer to the question 'How does it work?' and the latter concerns an answer to the question 'Who is responsible for the way it works?' As long as the notion of explicability is general and robust, there is a wide consensus on the significance of the explicability of ML algorithms.

[1] It includes two kinds of transparency: the failure and the judicial transparency. According to the failure transparency principle, if an AI system causes harm, it should be possible to ascertain why. According to the Judicial transparency, any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority (AI principle 2017).

[2] "We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyzes the world. Ethics and design and in hand" (Nadella 2016).

However, there are several unresolved issues on the principle of explicability: its moral status, what it is for, to whom AI needs to be explicable, and what kinds of explanation should be given to meet the principle. First, the normative status of the principle is not clearly stated (e.g., whether it is a categorical requirement or a recommendation). Second, regarding why the principle is needed, there is a controversy between the trust-based view and the autonomy-based view. While some argue that explicable AI is for human autonomy, many scholars claim that it is required to enhance users' trust in the system and its generated results. Third, there is no agreement on whether explicability ought to be directed toward decision-makers using the algorithms or toward decision-recipients affected by the algorithm-assisted decisions. Further, it is required to provide the criteria on which we assess the intelligibility and accountability of AI. Although it is sometimes recognized that different explanations need to be given to different stakeholders for different purposes, the problem of what kinds of explication are required (and how they meet the criteria) has yet to be thoroughly investigated.[3] For the reason of space, however, we confine our discussion to what-, why-, and whom-questions.

This paper aims to contribute to explicating the concept of explicability rather than providing a survey or description of its current usage in the relevant literature. By "explication," as defined in Carnap (1950, 1955), I mean a prescription of how we should characterize the concept in order to use it more productively. It involves replacing the unclear concept (the explicandum) with a clearer one (the explicatum) to serve the goals of the concept better. Our primary focus is to demonstrate the dual nature of explicability by addressing the why- and whom-questions. While this paper does not present entirely novel arguments for particular views over others, it adopts a methodological strategy of dissecting and rearranging the existing materials from the current literature on explicability. I contend that the seemingly plausible arguments failed to support the claim they are supposed to because they are ignorant of the dual nature of explicability. Some arguments are cogent in demonstrating one aspect of explicability but fail to account for another aspect, and *vice versa.* To a coherent set of answers to

[3] The question of what kinds of explanations are needed for different contexts will be addressed in a separate article.

the why and whom-questions, it is crucial to take into account the duality of explicability.

This paper will proceed as follows. In Section 2, I begin with setting the stage by examining how data science usually works in practice and suggest the principle of explicability as a prima facie duty to make ML algorithms explicable. In Section 3, I clarify the who(m)-question, addressing who is responsible for making algorithms explicable and to whom the explanation is owed. In Section 4, I delve into reasons for demanding explicability, highlighting the duality of autonomy and trustworthiness. In Section 5, it is claimed that the explication should be directed toward decision-makers as well as decision-recipients.

2. Setting the Stage

In our context under consideration, the primary function of ML algorithms is to offer prediction-based classification or ranking for specific purposes. It is often claimed, for example, that ML algorithms estimate the probability of a prisoner committing crimes again to determine their eligibility for release on parole or predict an applicant's credit score to classify her as qualified or unqualified for a bank loan (Pasquale, 2015; O'Neil, 2016). Of course, before the advent of ML algorithms, individuals have routinely been subject to classification: people have been ranked on credit score, deemed as qualified or disqualified for insurance, or accepted or rejected for various applications. Now, these classificatory tasks are being automated with ML algorithms.

How do automated or algorithm-assisted decision-making systems work in real-world cases? Roughly speaking, the decision-making processes consist of several steps that need to be iterated: defining the problem, data collection, building algorithm models, training and testing models with collected data, and application in real-world situations.[4] Think of a loan application system as an instance. First, the problem should be clearly defined:

[4] Roose (2023) introduces similar steps in explaining how AI technology actually works: 1) set a goal, 2) collect lots of data, 3) build your neural network, 4) train your neural network, 5) fine-tune your model, and 6) launch carefully.

classifying an applicant based on her credit and ability to repay the debt. Then, developers build models and collect data to train and test the model. If the tested model turns out to be successful, it will be applied in the decision-making of loan applications. Such algorithmic decision-making systems are to be used in a wide range of real-world situations.

Let me, in this paper, focus on situations where the use of ML algorithms is morally significant in that they involve direct assessment of personal capacities, characters, or properties that would affect the condition or quality of their lives (i.e., judicial sentencing, job interview, or predictive policing).[5] Given that the algorithms are used in this context, precisely what does the principle of explicability state? Although Floridi and colleagues (Floridi et al., 2018; Floridi and Cowls, 2019) put forth explicability as an overarching principle, they are not explicit on the normative status of the principle. As an ethical principle, it might be understood as a categorical imperative that ML algorithms be explicable or a recommendation that explicable algorithms be more desirable than inexplicable ones. Or, the principle might be interpreted as saying that it is morally wrong to use inexplicable AI in specific contexts.[6]

To clarify the normative status of the principle, we can take advantage of an analogy with biomedical ethics. In the meta-analysis of AI ethics guidelines, Floridi and Cowls (2019) found that AI ethics share four core principles of bioethics: beneficence, non-maleficence, autonomy, and justice

[5] I do not intend that those situations mentioned exhaust all the cases where using ML algorithms is ethically significant. Further, there are many other situations where algorithms have nothing to do with assessing personal capacities, characters, or properties (e.g., voice assistants), or no one is affected significantly by the algorithms in a moral sense (e.g., spam filters). However, the issue of accountability, the ethical aspect of explicability, arises when someone is affected by algorithm-assisted decision making.

[6] The failure transparency of Asilomar AI principles (2019) states that if an AI system causes harm, it should be possible to ascertain why. It sounds like a categorical imperative which should be obeyed when there is harm caused by AI systems. London (2019) takes the principle as a recommendation to prioritize explainability over predictive accuracy and criticizes it as unwarranted in the healthcare field.

(Beauchamp and Childress, 2012).[7] It is worth noticing that the four principles are usually regarded as prima facie duties. As they are not exceptionless requirements or imperatives, each principle needs to be weighed against other principles when applied to a concrete case and, when warranted, can be overruled by other principles. While they are instrumental in reflecting on moral problems and working toward an ethical resolution, they do not provide easy, ready-made solutions to particular cases. Likewise, we can take the principle of explicability as a prima facie duty for contemplating the moral problems of algorithmic decision-making systems, not merely an unqualified requirement for every application of AI algorithms. More specifically, I suggest the explicability principle as involving a prima facie duty to make ML algorithms explicable when applied in morally significant situations. As a prima facie duty, explicability does not necessarily take priority over others.[8] Nonetheless, the principle can be overruled only when there are good reasons justifying doing so.

Efforts to develop explicable ML systems can create tensions with other important values, including predictive accuracy, efficiency, privacy, and fairness. For the ethical development and deployment of ML systems, it is necessary to carefully manage the trade-offs among these competing values. This requires consideration of the specific context of deployment since no single solution is suitable for every situation. Suppose the ML algorithms are used in medical diagnosis, which often demonstrate superior predictive accuracy despite their highly complex and inscrutable structures. In this medical context, patient well-being is at stake, and diagnostic errors have severe consequences. Consequently, prioritizing accuracy over explicability can be ethically justified. Another type of trade-off involves the tension between explicability and efficiency. In situations demanding urgent decision-making (e.g., real-time financial trading or emergency response systems), rapid action is vital. In such demanding scenarios, the time delays

[7] It does not mean that they do offer a perfect translation. It only means that various ethical principles in different guidelines can be incorporated into four overarching principles with slightly different connotations.

[8] Of course, the remaining question is how to deal with a situation where different principles conflict. That is a subject to further discussion, depending on the context.

caused by generating detailed explanations may be practically untenable. Hence, in these contexts, operational efficiency can override the requirement for transparency.

Additional trade-offs become apparent when it comes to considering data privacy and fairness. For example, when the automated system is used to assess job applicants, providing exhaustive explanations might inadvertently compromise the applicant's privacy by revealing sensitive personal information. Moreover, full transparency can sometimes unintentionally undermine fairness. This can occur when simple, interpretable models rely on proxy variables closely correlated with protected characteristics (e.g., race or gender). Then, the principle of explicability may conflict with the demand for anti-discrimination. Paradoxically, a reduced degree of explicability in such situations might help to preserve privacy and promote fairness. Determining when the principle of explicability can be overridden requires careful, context-specific ethical deliberation. Decision-makers must assess the potential benefits we can get by improved accuracy or efficiency and the risks posed by diminishing privacy or fairness. Critically, the ethical legitimacy of any trade-off depends upon the acceptance or consent of the stakeholders affected by the decision-making systems.
Given that explicability is a prima facie duty, we must ask who holds the duty, to whom, and why.

3. Clarifying the Who and Whom Questions

Two questions, ‘Why does the principle of explicability matter?’ and ‘Who has the duty to make algorithms explicable to whom?’ are intimately related. As demonstrated subsequently, the answer to the first question informs the latter. In order to clarify the who(m)-questions, it is crucial to identify the key stakeholders involved in using automated decision-making systems. Among many stakeholders, five roles participate primarily in developing and utilizing ML algorithms: algorithm-developers, algorithm-users (and decision-makers), managing-users, decision-recipients, and regulators (cf. Arrieta et al., 2020; Langer et al., 2021). It appears that developers, algorithm-users, and decision-recipients are the most directly involved stakeholders with respect to the who/whom questions.

Algorithm-developers are those who are engaged in designing and developing ML algorithms. The developed and tested algorithms are used by algorithm-users. By algorithm-users, I mean those who make decisions using ML algorithms. Usually, algorithm-users are professionals (e.g., judges, police officers, and human resources managers) working in tandem with a system, or those who use the system make decisions that affect individuals. Decision-recipients (affected parties) are individuals or groups of individuals who are actually or potentially influenced by the system's decision. Typically, the three roles separately hold. Designers build the system on users' requests, and users make decisions about the applicants who are affected by the decision. For example, when algorithms are used in reviewing loan applications, developers design and develop the review algorithm, bank officers use it to judge whether applicants are eligible, and the decision made influences applicants.

In addition to the three main roles, other stakeholders are indirectly but significantly engaged in decision-making processes: managing-users and regulators. Managing-users (e.g., company, committee, etc.) decide to adopt an automated decision-making system with which algorithm users make decisions. Although they are not working with particular ML algorithms, they request developers to build a decision-making system with the desired specification and supervise the uses of the system. Thus, managing users are users of algorithms in an indirect sense. Finally, there are regulators (e.g., the government) who monitor and regulate the whole process and intervene in the cases where necessary. They are expected to establish moral and legal standards for the general use and development of automated systems. This class of stakeholders plays a unique role since they operate as a "watchdog" for both the systems themselves and the way in which they interact with the other stakeholders.

They are five different roles that people can play in decision-making processes. I prefer to talk about the roles rather than types of stakeholders because a single stakeholder can often play multiple roles simultaneously. For example, a CEO of an IT company who actively contribute to developing an algorithm for reviewing job applicants may also use it to make decisions, thus playing the roles of a developer, algorithm-user, and managing-user. Alternatively, one person (or a group of people) can be both a decision-maker and decision-recipient, where their decisions affect themselves. When we use a recommendation system for choosing books or movies (on Amazon or Netflix), we play the roles of algorithm-users and decision-recipients in that we make decisions affecting ourselves. In these cases, people decide whether to accept or reject the algorithm's suggestions or to use them for their actions. Still, the distinction of roles between designers, decision-makers, and decision-recipients holds.[9] Further, in our context of interest, where the issue of accountability is salient, the asymmetry between algorithm-users who make decisions with ML algorithms and decision-recipients who are affected by the decisions is assumed.

Given that the principle of explicability is about a prima facie duty to make ML algorithms explicable, we need to distinguish two who-questions: "Who has the duty?" and "To whom should the algorithms be explicable?" Let us call the former the who-question and the latter the whom-question. The answer to the former, I believe, is not very controversial. Anyone involved in designing, building, and using ML algorithms in ethically significant situations ought to contribute to making them explicable. First of all, algorithm-developers have a duty because they are the only ones who can make algorithms explicable (or inexplicable) in a direct sense. Nevertheless, the duty must be distributed among managing-users and regulators because they are supposed to contribute to developing and using the algorithms. Managing-users ought to have developers make ML algorithms explicable, while regulators must guarantee that the algorithms used in morally significant situations are explicable.

If we have a consensus on who has the duty, then the remaining question is: ML algorithms should be explicable to whom? One obvious response would be that ML algorithms should be explicable to every role of

[9] Medical AIs (e.g., imaging) are located in more complex situations. Developers and engineers build the medical AI system at the request of health professionals, who make use of the system to diagnose and treat diseases. Interestingly, decision-making in AI-assisted medicine is distributed among doctors and patients, although the final decision should be made by patients. With the aid of the ML algorithm, health professionals determine what the disease is and suggest a promising treatment. Patients accept (or reject) the treatment based on the doctor's suggestion (based on ML) and their own values and preferences.

stakeholders, including developers and regulators. Developers might think explicable AI is desirable because it helps to examine the limitations and errors and improve the performance of a system (e.g., debugging). To build a reliable system, however, it is unnecessary to make it explicable. Of course, there is a trivial sense that ML algorithms must be explicable to developers if they ought to be explicable to *anyone*. For the developers are those who are able to make it explicable and provide the explication to other stakeholders. From the regulator's perspective, explicability is needed to supervise and regulate the use of algorithms in morally significant decision-making. It does not necessarily mean the algorithms should be explicable to regulators in every morally loaded case. Instead, regulators might request the explicable AIs on behalf of decision-recipients and ordinary citizens because they are actual or potential parties concerned who are directly assessed and affected by the algorithmic decision-making systems.

The whom-question we are asking is more specific: to whom should the explanation be intelligibly given when ML algorithms are used in morally significant cases? According to a dominant view, what I call "the recipient-oriented view," algorithms should be explicable to individuals affected by the decision-making system's outputs (e.g., Grote and Berens, 2020; Kim and Routledge, 2021; Watcher et al., 2018). In other words, the developers ought to make the algorithms explicable so that the decision-makers using the algorithms can provide the decision-recipients with an explanation (e.g., "how the decision was produced"). Although most of the literature focuses on the explicability toward the decision-recipients, there is an alternative view, what we might call "the user-oriented view." For example, Robbins (2019) claims that explicability is not directed toward the person subject to the algorithm's decision, and algorithms should be explicable to those who make a decision using the algorithms. In other words, the algorithm-designers ought to make it explicable to decision-makers.

I suggest that the recipient-oriented view and the user-oriented view are not mutually exclusive but complementary to each other. In this regard, the problem with the two views lies in the assumption that explicability is directed to only one class of stakeholders involved in the decision-making processes, which has not been justified. The recipients-oriented view mistakenly assumes that the decision-recipients are those only who have the right to receive the explanation of the algorithm's decision. Robbins rightly highlights the significance of being explicable to decision-makers, which has been largely ignored by the proponents of the recipient-oriented view. However, he is also mistaken to maintain that explication is not directed toward the decision-recipients.[10] It is worth noticing that different stakeholders might want explicable AI for different purposes. An algorithm explicable to one stakeholder might be inexplicable to other stakeholders. This consideration motivates us to consider the possibility that while each view reflects a part of the whole process, explicability should be given to both decision-makers and decision-recipients. The duality of explicability with respect to the whom-question is intimately related to the dual goals of explicability. As the answer to the whom-question is informed by the way we answer the why question, we will discuss why the explicability of algorithms matters in the next section.

4. Why Algorithms Should Be Explicable

(1) Opacity and Request for Explicability

What is explicability for? Why should we adopt the explicability principle alongside other basic principles?[11] Regarding these questions, two main views comprise most of the literature: the trust-based and the autonomy-based views. Some scholars argue for the autonomy-based view, according to which the explicability helps humans working with algorithm models make their own informed decisions (e.g., Robbins, 2019). Others have advocated the so-called "trust-based view," according to which the explicability

[10] Robbins claims that "the person using the algorithm is the person that the explanation should be directed towards—not the person subject to the decision of the algorithm"(2019, p. 503). The reason why the claim is not justified will be examined in Section 6.

[11] From the managing user's perspective, it is necessary to comply with relevant regulations (e.g., GDPR's a right to explanation). However, such a compliance is merely a derivative reason while we are seeking for answers in a more fundamental level.

of algorithms is instrumental in enhancing the trust of decision-recipients in ML algorithms (e.g., Kim and Routledge, 2021).

Before we go further, the clarification of the trust-based view is in order. It is crucial to distinguish between trust and trustworthiness (Simon, 2013; McLeod, 2021). Note that people can trust someone who is untrustworthy and sometimes do not trust trustworthy persons. That is, the act of trust of the trustor in the trustee should not be conflated with the trustworthiness as a property possessed by the trustee. For business purposes, it would be advantageous to enhance people's trust in AI systems; however, the increase in trust itself is not constitutive of a moral obligation to make the system explicable. Instead, one can maintain that developers are morally obliged to make trustworthy algorithms[12], which requires explicability. In other words, we want people's trust in algorithms to be justified or warranted. In the following, when I refer to the trust-based view, I will mean the idea that explicability is required to ensure that decision-recipients' trust in algorithms is warranted. In the remainder of this section, I claim that the two views - the autonomy-based and the trust-based view - are not competing but complementary to each other.

The request for explicable AI seems to stem from the opaque nature of ML algorithms. In contrast to GOFAI, recent ML algorithms (i.e., deep neural net) are not transparent (for the forms of opacity, see Burrell, 2016).[13] It is widely held that the black-box nature of algorithms calls for

[12] The call for trustworthy AI abounds. For instance, European Commission (2019) has presented *Ethics Guidelines for Trustworthy AI* and OECD (2019) published *Recommendation on AI* which emphasize the "international co-operation for trustworthy AI."

[13] There are different senses in which ML algorithms are not transparent. Basically, the opacity of ML algorithms means that the recipients of algorithm's output have no understanding of why the decision-output has been made (and the inputs themselves are unknown or partially known). The opacity might mean intentional secrecy (no information open to recipients) or technical illiteracy (not easily understandable to non-experts). But there is a sense in which ML algorithms are fundamentally opaque even to engineers, which means "the mismatch between mathematical optimization in high-dimensionality character of ML and semantic interpretation demanded for human intelligibility" (Burrell, 2016). The fundamental sense of opacity is our focus in this paper.

the principle of explicability when used in morally loaded cases. If the outcomes of the black boxes have no moral significance, then explicability is not demanded.[14] For instance, no one would argue for a duty to make AlphaGo (Go-playing algorithm) explicable. Hence, explicability is called for when black boxes are used in morally significant situations. Our question is why it is so.

I claim that explicability is called for when opaque technologies threaten autonomy or trustworthiness. To put it in another way, the demand for explicability does not arise when 1) the technologies can be justifiably trusted and 2) users are able to control the uses of the technologies in that they judge whether (and/or when) they can use it or not based on the knowledge of its capabilities and limitations. For example, think of pills whose physiological mechanism on how it works is unknown. Still, if they are proven safe and effective, and we know when we can use them (and when we should not), there is no need to explain how it works. The problem is that the opaque nature of ML algorithms can threaten trustworthiness and autonomy (control to decide).

(2) Undermined Trustworthiness Calls for Explicability

Opacity undermines trustworthiness or warranted trust. Opaque algorithms make it difficult to judge whether decision-recipients' trust in them is warranted. Some scholars think that reliability is better for conferring trust than explanation.[15] For example, London (2019) argues that prioritizing explicability over reliability is misleading and criticizes the call for an explanation as unwarranted. Indeed, there are many cases where new

[14] Robbins (2019) made a similar point.

[15] Even when ML algorithms have been shown to increase the accuracy of diagnosis, a lack of explanation for the diagnosis can lower doctors' trust in the algorithms (Ribeiro, Singh, and Guestrin, 2016; Creel, 2020). However, it is unclear whether people will trust AI more when given some explanation. Recently, there is a growing body of empirical work on how providing an explanation affects people's trust in ML and its decision (e.g., Lu et al., 2019). But they usually focus on types of explanation (i.e., decision tree or diagram) to effectively raise people's trust. Here we are not concerned with mere increase of trust but with the warrant of the trust or trustworthiness.

technology can be trusted when it works well by reliably generating the desired output. Pills might be justifiably 'trusted' if their functionality is well-proven. Thus, we can ask whether explicability is required when ML algorithms turn out to be reliable. London and others would say that reliability suffices while explicability is not required.

It is widely held that we should (or do) pursue AI technologies that are not merely reliable but also trustworthy (European Commission, 2019; OECD, 2019). One of the consensuses of philosophical literature on trust and trustworthiness is that trust is not mere reliance, although it is a kind of reliance. As Baier (1986, p. 235) succinctly put it, "trusting can be betrayed, or at least let down, and not just disappointed." While reliance or reliability can be defined in terms of rational expectation, trust or trustworthiness is often defined in terms of normative or moral expectations of trustors. Unlike mere reliance, trust includes a normative expectation whose violations induce to betrayal, not merely disappointment. Trustworthiness can be defined by the property of a person who can be justifiably trusted. As trust can be betrayed, it is valuable only when directed to trustworthy persons.

To establish trustworthiness, two requirements must be met: competency and responsiveness. As articulated by Jones (2012), a person is trustworthy with respect to the trustor in domain of interaction D if and only if "she is competent with respect to the domain, and she would take the fact that [the trustor] is counting on her (...) to be a compelling reason for acting as counted on" (pp. 70–71). Competency refers to a person's ability to fulfill the trustor's expectation with respect to the tasks she is supposed to do. Responsiveness, on the other hand, involves taking into account and responding to the trustor's interests and values. While a technological artifact can be deemed competent if it can fulfill the user's expectations regarding its task, it cannot meet the responsiveness condition.

To apply the notion of trust and trustworthiness to AI technologies, we need to consider a socio-technical system incorporating technological artifacts, human agents, and institutional structure (Nickel et al., 2010, Rieder et al., 2021). Under this interpretation, the trustee is not a technological artifact but a network of technical objects and human agents. Accordingly, when someone 'trusts' a particular technology, she trusts a socio-technical system that includes designers, operators, and other stakeholders interacting within a regulatory or legal framework. In this regard, pills or other technical products (as a component of socio-technical systems) can be regarded as trustworthy to the extent that the human agents involved in developing and operating them are trustworthy, and the evaluative system for testing them is validated (e.g., industrial standards or certificates). It is the socio-technological system that can be responsive to the user's interests and values.

It is worth noticing that reliability in the sense of well-functioning does not guarantee the trustworthiness of technology in cases where due process is crucial. For an ML algorithm to be trustworthy, it must be both competent and responsive. ML algorithms count as competent if they can meet the trustor's expectations in the domain of decision-making. However, when algorithms are opaque, it is difficult to satisfy the expectations because people expect algorithmic decision-making to be fair, unbiased, and non-discriminatory. To fulfill the trustor's expectation, the fairness of the decision-making processes must be ensured, necessitating explicability. How, then, can a socio-technical system, which incorporates ML algorithms as a component, meet the responsiveness requirement? It is not only essential for the human agents involved to be trustworthy, but the evaluative system for assessing the ML algorithms should be validated. The algorithms should be explicable to ensure that the evaluative system is responsive to the trustor's values and interests.

If the algorithm is opaque, it is impossible to discern what bases the decision made. Consequently, we lack an understanding of why it generates its outcomes and whether they are acceptable and justifiable. Therefore, the decisions generated from the opaque processes do not guarantee trustworthiness. To build trustworthy AI, it is required to make the ML algorithms explicable because explicability is a crucial means to examine the fair and justifiable application of AI.[16]

[16] While we have been focusing on the warranted trust of decision-recipients in algorithms, the trust of algorithm users (decision-makers) matters too. They are those who use the algorithm in morally significant cases. If they do not have warranted trust in it, they are not likely to use it. Furthermore, as they make an impact on the recipients of the decision, they must be held accountable to the recipients. Thus, decision-makers also care about the trustworthiness of ML algorithms.

(3) Undermined Autonomy Calls for Explicability

Opacity undermines autonomy. The principle of autonomy in AI ethics concerns a balance between human-led and machine-led decision-making (The Montreal Declaration for Responsible AI, 2017) or “between the decision-making power we retain for ourselves and that which we delegate to artificial agents” (Floridi and Cowls, 2019). Humans have to decide “whether to delegate decisions to AI systems, to accomplish human chosen objectives” (Asilomar AI Principles, 2017). According to Floridi and others, the principle of autonomy states that “humans should retain the power to decide which decisions to take: exercising the freedom to choose where necessary and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making” (Floridi and Cowls, 2019).

When it comes to opaque algorithms, humans may lose control to decide. If we lack a good grasp of how the model works, it becomes very difficult to consider good reasons to choose whether to use it (i.e., whether to accept or reject the decision suggested by the system). Consequently, the autonomy of decision-makers using ML algorithms is compromised. In line with this spirit, Robbins (2019) asserts that the principle of explicability is primarily for maintaining meaningful human control over ML algorithms. Here, the expression “meaningful human control” originates from the discussion around autonomous weapon systems and refers to the control human operators have over algorithmic uses of weapons. By generalizing this notion, Robbins (2019) adopts a specific conception of meaningful human control as “giving humans the ability to accept, disregard, challenge or overrule an AI algorithm’s decision” (p. 496).

Suppose that an ML algorithm for medical diagnosis predicts that the symptom indicates skin cancer while giving no explanation. Further, suppose that your doctor does not understand the functioning of the model and when to rely on it (or when not to). Even if the algorithm is more accurate in diagnosing diseases than humans, the doctor might lack the control to decide. Doctors, as algorithm-users, have responsibility for their use to the patients. If the algorithm is not explicable to the doctor who accepts the model’s decision that it is cancer, how can she be held responsible for her patients? As algorithm users make a decision and hold accountable for the decision-recipients who are affected, they need to know how the outputs are generated.

While Robbins correctly points out that explicability is for meaningful human control over ML algorithms, his assertion that autonomy is the sole objective of making ML algorithms explicable is misguided. It is true that explicability is instrumental to achieving autonomy, but this is only part of the whole story. As previously demonstrated, trustworthiness is another crucial goal of the explicability principle. Trustworthiness and autonomy are not mutually exclusive goals, but they are complementary ones. In summary, the principle of explicability has dual objectives: trustworthiness and autonomy.

5. The Dual Nature of Explicability

The remaining question is to whom the explication is owed. Given the dual goals of explicability, I contend that explication should be directed toward both decision-recipients for their warranted trust in ML algorithms and decision-makers for their informed decisions.

First, concerning the objective of trustworthiness, the whom-question can be rephrased as follows: who are the trustors to whom algorithms should be trustworthy? The decision-recipients affected by algorithmic decision-making are the trustors in question. Consequently, the recipient-oriented view on the whom-question is well aligned with the trust-based view on the goal of explicability. ML algorithms are used to assess the personal capacities or characters of decision-recipients. For a decision-making system utilizing ML algorithms to be considered trustworthy, it must fulfill the recipient's expectations and remain responsive to the decision-recipients by operating on behalf of their interests and values. When ML algorithms are used in morally significant situations, it is essential to guarantee that the decision-recipients' trust in the algorithms is warranted. Therefore, there is a good sense in which explication should be given to the decision-recipients.

Second, when it comes to retaining autonomous decisions, ML algorithms should be explicable to the decision-makers employing them. In order to make informed and responsible decisions, algorithm-users should be able to decide whether to accept, modify, or reject the recommendations generated by the algorithms. To achieve this, they must have an understanding of algorithms' functionality, capabilities, and limitations. Providing the explication of how they work enables decision-makers to make more informed and autonomous decisions, particularly when their decisions have consequences of ethical significance. Thus, the autonomy-based view informs us that explication should be directed toward the decision-makers.

Before we conclude the duality of explicability, a critical examination of Robbins' criticism of the recipient-oriented view is in order. Without any justification, Robbins (2019) assumes that the explicability principle has only one goal - maintaining meaningful human control over algorithmic decisions. He posits that with an explanation of the algorithm's decision, human beings can retain their control to decide. Although he is aware of "the ethical issues of ensuring that the outputs of algorithms are not made based upon ethically problematic or irrelevant considerations" (p. 501), he regards it as an aspect of meaningful human control. Consequently, he maintains that "an explanation of the algorithm's decision can allow for someone to accept, disregard, challenge, or overrule the rejection. This gives meaningful control of the decision to human beings"(ibid.). It would make sense if the "human beings" refer to the decision-recipients. However, Robbins contends that the explanation should be given to the decision-makers instead. While he concedes that a decision-recipient "subject to the algorithm's outputs may be interested to know the explanation," he asserts that her interest "does not establish meaningful human control over the algorithm's output" (pp. 502–503).

Robbins claims that explanations are not helpful to the decision-recipients. For example, in the case of a loan application rejected by an algorithm, the explanation of rejection may include a high debt-to-income ratio of the applicant. Without relevant domain knowledge, the applicant would be unable to assess whether the debt-to-income ratio considered was justifiable ground for rejection. Thus, Robbins concludes that the explanation fails to establish meaningful human control, and as a result, explanations need not be directed toward decision-recipients. However, this line of thought is flawed. First, as we have seen, the decision-maker's control to decide is not the only purpose of explicability. Second, our claim is not that explanations given to the decision-recipients are always sufficient for them to judge whether the algorithmic decisions are based on unethical or problematic considerations. For such judgments, decision-recipients need to have relevant information and background knowledge. Nonetheless, it does not undermine our claim that explicability is necessary for decision-recipients.[17] Most of the literature on explicability has mistakenly assumed that explication should be provided to just one stakeholder role for one and only purpose. By rejecting this unjustified assumption, we can embrace the dual nature of explicability. The duality of explicability is two-folded: first, explication serves the dual purposes of trustworthiness and autonomy; second, explanations should be directed to decision-recipients as well as decision-makers. Moreover, these two dimensions seem to overlap coherently, as explanations are needed to warrant decision-recipient's trust in ML algorithms and facilitate informed decisions by decision-makers.

6. Conclusion

Although many authors have made significant contributions to answering why ML algorithms should be explicable and to whom, they have failed to construct a coherent, systematic framework for conceptualizing and implementing explicability. In this paper, I have attempted to fit the pieces of the puzzle together. First, I argued that explicability is instrumental in enhancing more fundamental values like autonomy and trustworthiness. Second, I highlighted the dual nature of explicability, particularly aligning the answers to why- and whom-questions along the two dimensions. Explicability is directed toward both decision-makers for their autonomous and informed decisions and decision-recipients for their warranted trust in the algorithms. The next question we must address is how different explicatory strategies should be employed for different stakeholders, depending on the purposes of explicability. I suspect that intelligibility, the epistemic sense of explicability, can be met by pursuing an objectual understanding of how

[17] Robbins (2019), in footnote 12, mentioned the goal of actionable recourse, "the ability to contest incorrect decisions or to understand what could be changed in order for the data subject to achieve a more desirable result" but failed to consider further implications (p. 503).

the models work (for algorithm-users) or a causal/counterfactual explanation of why a particular decision was made (for decision-recipients). However, that is a task I aim to tackle in another article.

Acknowledgements

Earlier drafts of this article were presented at the lunchtime talk of the Center for Philosophy of Science at the University of Pittsburgh, the AI Policy Forum of the Korean Association of AI and Law, and the 10th Biennial Meeting of the Asia-Pacific Philosophy of Science Association. I am especially grateful to Edouard Machery, Collin Allen, Haksoo Ko, Gunoo Kim, and the participants for their valuable comments.

Funding

This research was supported by the Global-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301976) and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00343989, Enhancing the Ethics of Data Characteristics and Generation AI Models for Social and Ethical Learning).

References

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

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

Montreal Declaration for a Responsible Development of Artificial Intelligence. 2017. Announced at the conclusion of the Forum on the Socially Responsible Development of AI. https://www.montrealdeclaration-responsibleai.com/the-declaration

Baier, Annette. 1986. Trust and Antitrust. *Ethics,* 96(2), 231–260.

Beauchamp, Tom and Childress, James. 2012. *Principles of Biomedical Ethics* (7th ed). New York: Oxford University Press. https://doi.org/10.1093/occmed/kqu158

Binns, Rueben. 2018. Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Burrell, Jenna. 2016. How the Machine' Thinks:' Understanding Opacity in Machine Learning Algorithms. *Big Data & Society.* https://doi.org/10.1177/2053951715622512

Carnap, Rudolph. 1950. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

Carnap, Rudolph. 1955. Meaning and Synonymy in Natural Languages. *Philosophical Studies*, 7, 33–47.

Creel, Kathleen. 2020. Transparency in Complex Computational Systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

European Commission. 2019. *Ethics Guidelines for Trustworthy AI.* [online] Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28: 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review.* https://doi.org/10.1162/99608f92.8cd550d1

Grote, Thomas, and Philipp Berens (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205. https://doi.org/10.1136/medethics-2019-105586

Goodman, Bryce, and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine*, *38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Jones, Karen. 2012. Trustworthiness. *Ethics*, 123(1), 61–85. https://doi.org/10.1086/667838

Kim, Tae Wan, and Bryan R. Routledge. 2021. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly*, 1–28. https://doi.org/10.1017/beq.2021.3

Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296: 103473. https://doi.org/10.1016/j.artint.2021.103473.

McLeod, Carolyn. 2021. Trust. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2021/entries/trust/>.

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

Lu, Joy, Dokyun (DK) Lee, Tae Wan Kim, and David Danks. 2019. Good Explanation for Algorithmic Transparency. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3503603.

Nadella, Satya. 2016. Microsoft's CEO explores how humans and AI Can solve society's challenges— together. Slate. https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

Nickel, Philip J., Maarten Franssen, and Peter Kroes. 2010. Can We Make Sense of the Notion of Trustworthy Technology? *Knowledge, Technology & Policy* 23: 429–444. https://doi.org/10.1007/s12130-010-9124-6.

OECD (Organisation for Economic Co-operation and Development). 2019. *Recommendation of the Council on Artificial Intelligence*. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Penguin Random House.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, MA: Harvard University Press.

Rieder, Gernot, Judith Simon, and Pak-Hang Wong. 2021. Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. In M. Pelillo and T. Scantamburlo (eds.) *Machines We Trust: Perspectives on Dependable AI*, The MIT Press. https://doi.org/10.7551/mitpress/12186.003.0007

Ribeiro, Marco Tulio, Sammer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins, Scott. 2019. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Roose, Kevin. 2023. How Does ChatGPT Really Work? *New York Times* (March 28, 2023) https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html

Simon, Judith. 2013. Trust. In: D. Pritchard (Ed.), *Oxford Bibliographies in Philosophy*. New York: Oxford University Press. Available online at: https://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577- 0157.xml