The Conflict Between Explainable and Accountable Decision-Making Algorithms

Gabriel Lima
gabriel.lima@kaist.ac.kr
School of Computing, KAIST & Data Science Group, IBS
Republic of Korea

Nina Grgić-Hlača
nghlaca@mpi-sws.org
Max Planck Institute for Software Systems & Max Planck Institute for Research on Collective Goods
Germany

Jin Keun Jeong
jkjeong@kangwon.ac.kr
Kangwon National University
Republic of Korea

Meeyoung Cha
mcha@ibs.re.kr
Data Science Group, IBS & School of Computing, KAIST
Republic of Korea

ABSTRACT

Decision-making algorithms are being used in important decisions, such as who should be enrolled in health care programs and be hired. Even though these systems are currently deployed in high-stakes scenarios, many of them cannot explain their decisions. This limitation has prompted the Explainable Artificial Intelligence (XAI) initiative, which aims to make algorithms explainable to comply with legal requirements, promote trust, and maintain accountability. This paper questions whether and to what extent explainability can help solve the responsibility issues posed by autonomous AI systems. We suggest that XAI systems that provide post-hoc explanations could be seen as blameworthy agents, obscuring the responsibility of developers in the decision-making process. Furthermore, we argue that XAI could result in incorrect attributions of responsibility to vulnerable stakeholders, such as those who are subjected to algorithmic decisions (i.e., patients), due to a misguided perception that they have control over explainable algorithms. This conflict between explainability and accountability can be exacerbated if designers choose to use algorithms and patients as moral and legal scapegoats. We conclude with a set of recommendations for how to approach this tension in the socio-technical process of algorithmic decision-making and a defense of hard regulation to prevent designers from escaping responsibility.

CCS CONCEPTS

• **Applied computing** → *Law*; *Psychology*; • **Social and professional topics** → *Governmental regulations*; • **Computing methodologies** → *Philosophical/theoretical foundations of artificial intelligence.*

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3534628

KEYWORDS

Responsibility, Accountability, Explainability, Artificial Intelligence, AI, Decision-Making, Algorithms, Blame, Designers, Patients, Users

ACM Reference Format:
Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3534628

1 INTRODUCTION

Artificial Intelligence (AI) is now used in a wide range of situations, from low-stakes scenarios like entertainment [90] to high-stakes life-or-death decisions like selecting who should be prioritized for medical help [75]. Extensive research has inquired whether algorithmic decision-making has negative implications for society. Studies have observed, for instance, algorithmic bail decisions to be racially biased [2], discussed how AI systems[1] used for hiring decisions could embed biases [8], and found online advertisement to discriminate against women [27].

A major problem with most decision-making algorithms is their opacity. Most algorithms are black boxes that do not offer explanations for their decisions, recommendations, or processing [76]. This limitation has been a central motivation for developing Explainable Artificial Intelligence (XAI), which proposes to make algorithms explainable by "making [their] functioning clear and easy to understand" [3]. In the context of algorithmic decision-making, XAI creates models whose behavior can be easily understood (i.e., those that are transparent) or that can explain their behavior after a decision (e.g., by providing post-hoc explanations). Calls for XAI have become widespread in industry, academia, and policymaking [48].

The XAI field aims to create systems that facilitate attributing responsibility to human agents involved in their development and deployment. Assigning responsibility for algorithmic decisions has been widely debated to be a difficult task due to the existence of a responsibility gap [4, 69]. As argued by Robbins [81], explainable systems would maintain meaningful human control, allowing

[1] We use AI systems and decision-making algorithms interchangeably. This is in line with the Explainable AI (XAI) literature. We note that most decision-making algorithms would be labeled as AI systems regardless of their complexity or autonomy.

responsibility to be traced back to designers,[2] users, and patients (i.e., those subjected to algorithmic decision-making). That is not to say that XAI is only put forward to deal with responsibility issues. Explanations can, for instance, also be used to comply with legal requirements, promote trust in decision-making algorithms, and assess their accuracy [57].

We argue in this paper that XAI is not a panacea to the plethora of responsibility issues that autonomous decision-making algorithms entail. We focus our discussion on AI systems that are designed to make consequential decisions and can provide explanations afterwards, i.e., algorithms that are explainable in a post-hoc manner. While we agree that explainable systems are necessary for the responsible deployment of algorithmic decision-making, we show how XAI's post-hoc explanations may be at odds with the public's understanding of AI systems' agency and blameworthiness. Furthermore, we discuss how those who are subjected to algorithmic decisions (i.e., patients) may be perceived as having meaningful human control over XAI systems and illustrate how this impression is false and does not translate to true empowerment over algorithms.

Considering blame as a response to the reasons upon which an agent has acted [86], post-hoc explainable algorithms may be perceived as actors that can explain the reasons behind their decisions and thus as blameworthy. Explainable AI systems may also be viewed as more capable and intentional than their opaque counterparts, resulting in higher levels of blame [63, 64]. This impression obscures the responsibility of human agents in algorithmic decision-making and shifts laypeople's moral judgments towards machines, potentially influencing policymakers and hindering the adoption of beneficial AI technologies [11, 19].

Motivated by the concern that developers could launder their agency for the deployment of autonomous systems [82] and implement superficial ethical measures to avoid regulation [38], we show how they could use XAI to create a false sense of understanding and control for patients. We illustrate this misleading impression with research showing that algorithmic explanations are often nonsensical and leave individuals with no real control [83]. XAI systems can also be used to deceive patients, even those trained in AI-related areas [33], creating moral and legal scapegoats.

By illustrating how the responsibility for explainable systems might be blurred, we contribute to the literature by analyzing the responsibility gap posed by autonomous systems with a novel and critical perspective on the XAI field. We conclude with a call for interpretable systems, which would emphasize developers' responsibilities throughout the development and deployment of decision-making algorithms. Finally, we discuss how current regulatory approaches fail to address the conflict between explainability and responsibility and offer potential solutions.

2 BACKGROUND

2.1 Explainable AI

Algorithms are used to assist human judges in bail decisions [2], decide which patients should be prioritized for medical assistance [75], evaluate job applicants [8], and in many other applications. Algorithmic decision-making has become widespread in society, and much research has been devoted to understanding its benefits and drawbacks. A common criticism of most models used for making these life-changing decisions is that they are inscrutable black boxes [87]. Users, patients, and even designers do not understand how algorithms make decisions, making it impossible to backtrack their decision-making process.

The widespread use of consequential decision-making algorithms has made understanding how they work necessary. Explainable Artificial Intelligence (XAI), a field committed to increasing people's understanding of decision-making algorithms, arose from this need. As defined by Arrieta et al. [3], an explainable system is "one that produces details or reasons to make its functioning clear or easy to understand" for a specific audience, be it users, designers, patients, or policymakers.

Efforts to develop explainable systems are often categorized into two groups: they either propose *transparent* models or build techniques to assist black box models to explain their behavior after a decision (referred to as *post-hoc explainability)* [61]. Algorithms are transparent when a human can simulate its functioning (i.e., the model is simulatable), explain each part of the model (decomposable), and follow its decision-making process. Post-hoc explainability, on the other hand, is supported by strategies and models that explain decisions for any given input.

A linear regression is a typical example of a transparent model. Variables can be human-readable, and people can simulate the model if it is not unnecessarily complex. There are countless examples of post-hoc explainability, including techniques for simplifying an algorithm's specific decision (e.g., LIME [80]), creating human-understandable visualizations (e.g., [88]), and presenting counterfactual examples (e.g., [104]). Post-hoc explanations present extra information, such as which feature of an input had the greatest impact on the final decision or similar examples that might have resulted in a different determination (i.e., a counterfactual). This paper focuses on post-hoc explanations and how they may conflict with accountability.

2.2 Widespread Calls for Explainable AI

In a study of the global guidelines referring to the ethics of AI, Jobin et al. [48] found explainability (and similar notions like transparency) to be the most prominent principle across the efforts to promote the responsible development and deployment of AI. Although not restricted to algorithmic decision-making, several guidelines proposed explainability for this specific context [25, 39]. Industry leaders have also called for research in the field of XAI. For instance, Microsoft's CEO Satya Nadella defended the use of transparent systems, "AI must be transparent. [...] People should have an understanding of how the technology sees and analyzes the world" [73]. Researchers have also observed an exponential increase in the number of articles addressing XAI being published in academic venues in the last few years [3], demonstrating the role of academia in the development of explainable algorithms.

Scholars have argued for explainable systems based on a variety of premises. Felzmann et al. [1] justified the research agenda to

[2] We refer to developers and designers as the *collective* agents that encompass programmers, executives, and any other entity involved in the design and development of decision-making algorithms. More specifically, we turn our attention to corporations that may develop decision-making algorithms.

promote the acceptance of AI systems. Mittelstadt et al. [72] defended explainability as a form of promoting trust and verification in algorithmic decision-making. Burrell [18] argued in favor of explainable systems as a form of assessing fairness. Systems that are understandable for stakeholders could also contribute to their privacy [3].

The development of XAI has also been motivated by legal requirements [10]. Existing laws, such as those granting consumers access to their credit score, require algorithms to explain their decisions [87]. The "right to explanation" included in the European General Data Protection Regulation (GDPR) is a good example of how politicians could enforce explainability [42]. While interpretations of the law are yet to take place in courts, GDPR seems to entitle patients the right to ask for explanations concerning the logic used by algorithms that make consequential decisions [87].

Although the introduction to XAI presented above portrays it as a panacea to the opacity posed by algorithmic decision-making, the field has also been subjected to serious criticism. Scholars have raised concerns about how explainable systems may give designers unwarranted control over the information that patients and users receive [9]. Kasirzadeh and Smart [52] criticized counterfactuals as an appropriate approach to XAI as it might require incoherent social theories, e.g., by disregarding that some social categories are immutable. De Laat [28] argued that full transparency should be avoided due to potential negative consequences, such as loss of privacy. Robbins [81] showed that requiring explainability of all AI systems is misguided and suggested that a principle of explainability should be connected to specific decisions rather than the technology.

Building upon this previous work, our novel critique of XAI relies on its conflict with one issue it seeks to solve: the responsibility gap posed by autonomous systems. Scholars defend that XAI contributes to the report of an algorithm's negative impacts and auditability [3]. Explainability has often been framed as a necessary condition for accountable AI systems (e.g., [70]). Similarly, Robbins [81] argued that explainable systems are primarily for maintaining meaningful human control and responsibility. Before delving deeper into how XAI could clash with accountability, we introduce the notion of the responsibility gap and how scholars have addressed it below.

2.3 Responsibility Gaps

Algorithms decide who is automatically enrolled in health care programs [75] and determine which job opportunities are shown online to job seekers [27]. If these systems are found to discriminate against a certain social group, who should be held responsible for any harm caused by this decision? One may claim that the designers should have foreseen this risk during the development phase, thereby making them responsible. Another possibility is to hold those who use these systems accountable for their decision to use algorithms in the real world. Because these algorithms examine past data to make decisions, persons who are subjected to algorithmic decisions (i.e., patients) may be viewed as important actors who influenced the decision. The dilemma of which of these entities is an appropriate subject of responsibility arises for autonomous and adaptive algorithms (e.g., AI systems), creating a responsibility gap.

The debate surrounding the responsibility gap was initiated by Matthias [69], who argued that autonomous and self-learning machines threaten the necessary conditions for holding someone responsible. The degree of control that operators, users, and manufacturers have over a machine's behavior is challenged by high levels of autonomy. Any attempts to predict the behavior of systems built to constantly learn and adapt to new contexts are limited, conflicting with epistemic requirements for accountability. This lack of control and knowledge creates a responsibility gap, under which no one is a suitable subject of responsibility. This gap refers to various aspects of responsibility, ranging from who is to blame (e.g., blameworthiness) to who should be held accountable for algorithmic harm to the public (e.g., public accountability) [30].

2.4 Bridging the Responsibility Gap

How may society bridge the responsibility gap? Some scholars argue that algorithms, no matter how intricate, autonomous, or self-learning, are just human tools [15]. They contend that an algorithm's behavior should be understood as a collection of its users' and designers' decisions and intentionality rather than as intentional actions per se [49]. These viewpoints maintain human agents as the subjects of responsibility for legal coherence [16].

Coeckelbergh [24] takes a different approach and proposes to ground responsibility on the patients of algorithms' actions to circumvent the responsibility gap. This approach proposes that those who use and develop AI systems should take on forward-looking responsibilities by focusing on how algorithms could harm those subjected to it. Champagne and Tonkens [21] present a similar proposal in the context of automated warfare, arguing that a "person of sufficiently high standing could accept responsibility for the actions of autonomous robotic devices." This perspective appears to be the approach taken by some self-driving car manufacturers, who have pledged to bear responsibility for any accidents caused by their machines [55].

Another approach for bridging the responsibility gap is to hold AI systems accountable. Stahl [93], for instance, claims that machines could be held (quasi-)responsible, fulfilling several social goals derived from moral responsibility. Coeckelbergh [23] discusses how AI systems could be held "virtually" responsible to the extent that they appear to be morally responsible. These views are related to the proposal to extend legal personhood and responsibility to AI systems [43, 95]; yet, such a proposal has encountered much opposition [16, 91].

The proposal to hold algorithmic systems responsible has been controversial. Punishing an AI system, for instance, would be meaningless because it is incapable of suffering or carrying culpability [26, 92]. Others have argued that holding AI systems accountable requires metaphysical properties that existing actors lack, such as consciousness [46] and sentience [94].

A less controversial view proposes to view AI systems as human-AI collaborations. For instance, Nyholm [74] argues that machines' agency should be viewed as a collaborative effort between them and human actors, in which the latter "initiate, supervise, and manage the agency" of the former. Such a viewpoint argues that humans should remain the locus of responsibility. Another comparable view defends a form of joint responsibility, in which humans and autonomous algorithmic systems share responsibility [44, 45].

3 THE CONFLICT BETWEEN EXPLAINABILITY AND RESPONSIBILITY

In this paper, we focus on XAI systems that make consequential decisions, such as those used in the medical [75], hiring [105], and financial [54] domains. Despite the fact that most AI systems are designed to work alongside human decision-makers, algorithms may nonetheless operate as decision-makers in practice. For instance, AI determines which job applications should be later evaluated by human employers, having a direct impact on employment prospects [105]. These AI systems are often marketed as decision-making tools, influencing how they are used in the real world [54]. The responsibility gap becomes significant in these scenarios because human decision-making is undermined or even dismissed, blurring human responsibility.

We argue that XAI poses a unique challenge and may contribute to the responsibility gap (see Figure 1 for a summary of the main arguments). First, we show how AI systems that provide post-hoc explanations can be viewed as blameworthy agents, obscuring human responsibility. Second, we argue that explainable algorithms create the false impression that patients have meaningful human control, leading to incorrect attributions of responsibility. Third, we look at how designers' control over XAI systems allows them to use algorithms and patients as scapegoats, escaping responsibility.

3.1 Explainable Systems as Responsible Agents

What does it mean to blame someone? Scanlon [86] has proposed a distinct interpretation of blame, in that judgments of blameworthiness are based on one's assessment of others' reasons behind their intentions and attitudes that go against the standards of their relationship, i.e., their "social contract." To blame someone is to respond to this impairment by modifying one's views on their relationship with the blamee.

Based on Scanlon's interpretation, we suggest that decision-making algorithms may be deemed blameworthy if they appear to provide the reasons for their decisions. When someone requests an explanation for an algorithmic decision, they expect to hear the reasons for that determination [31]. As a result, AI systems that can provide post-hoc explanations may be deemed blameworthy if their reasons go against what is expected from them. It is worth noting that one of the issues covered by the responsibility gap is blameworthiness [30].

There is evidence that people blame AI systems and algorithms when they cause harm, regardless of their explainability. For instance, people blame algorithms when they make life-or-death decisions about who should live or die in a moral dilemma [66]. Another study found that AI systems described as autonomous are blamed to a similar extent to human agents [41]. Research has also shown similar results in several scenarios, ranging from medicine [59] to autonomous vehicles [40]. In the context of algorithmic decision-making, AI systems making bail decisions are blamed similarly to human judges [60]. In this paper, we do not argue that explainability alone affects blameworthiness. Instead, we argue that explainable systems may be subjected to higher levels of blame, conflicting with its aim to trace responsibility back to human actors.

One possible criticism of the argument above is that blameworthiness does not encompass all notions of moral responsibility. Moral philosophers have highlighted the pluralistic component of moral responsibility (e.g., [98]). For instance, Shoemaker [89] differentiates between three different notions of moral responsibility: attributability, answerability, and accountability. While attributability does not seem to apply to AI systems because they rely on evaluations of an agent's character, the latter two notions pose the same contradiction we defended above. First, Shoemaker defines responsibility-as-accountability similarly to Scanlon's blameworthiness, which we discussed above. Second, Shoemaker poses that to be answerable relies on an agent's ability to cite the reasons upon which their actions are grounded. Hence, AI systems that can explain the reasons behind their decisions could also be perceived as morally responsible with respect to answerability. We do not claim that XAI systems will be perceived as responsible according to all definitions of moral responsibility; nevertheless, this perception could obscure the moral responsibility of human agents in algorithmic decision-making.

Empirical evidence also supports the role of reasons in assigning blame and moral responsibility. Blame can be mitigated if the agent has justifiable reasons; in contrast, blame can be increased if the agent bases their actions on improper or immoral reasons [64]. The perception of an agent acting upon reasons could lead to higher perceived intentionality. Beliefs and desires, which can be given as reasons for an action, are crucial components of the folk conception of intentionality [65]. Explainable algorithms' decisions could be perceived as intentional, as suggested by prior research showing that humans adopt an intentional stance towards non-human entities [77] and robots [68]. Most crucially to our argument, prior work has shown the pivotal role of intentionality in judgments of blameworthiness [63]; if XAI systems are perceived as intentional, they are likely to be deemed blameworthy.

Explainable AI systems could also be perceived as blameworthy without attributions of intentionality. Malle et al.'s theory of blame [64] proposes that agents who are perceived to have the capacity to prevent or contribute to an event are also attributed substantial levels of blame. People view XAI systems are largely capable, as evidenced by research showing that individuals over-rely on algorithmic recommendations [7, 47], partly because they overtrust the explanations provided by these systems [6]. Explainable systems give the impression that they can make good decisions, and people blindly follow their decisions. This sense of capacity can also result in heightened blame attributions.

All of the research presented thus far suggests that decision-making algorithms could be seen as morally responsible if they can explain their decisions. Therefore, people may hold explainable algorithms responsible for their decisions, even though they may not be appropriate subjects of responsibility [26, 46, 92, 94]. XAI aims to trace responsibility back to human agents, but it may instead shift responsibility to the algorithms it aims to explain.

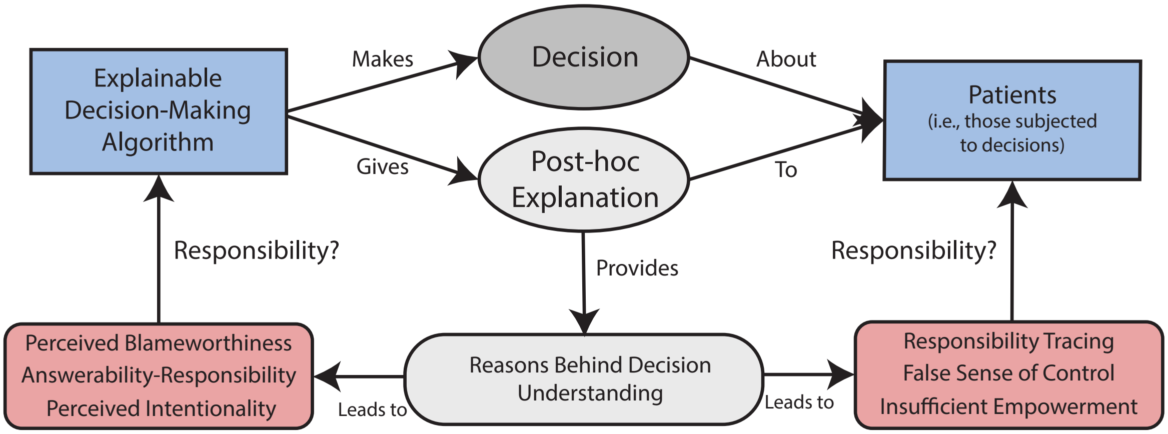

Figure 1: Our main points are summarized in this diagram. Decision-making algorithms that provide post-hoc explanations can be viewed as blameworthy agents that should be held accountable for their decisions; we explain how this perception can be explained by increased attributions of intentionality and capacity. Explainable systems can also result in incorrect attributions of responsibility to those subjected to algorithmic decisions (i.e., patients). Explainable systems give the impression of confidence and empowerment, implying that patients should bear some responsibility. This notion is incorrect and could be exploited by designers attempting to escape responsibility for algorithmic decision-making.

3.2 Patients as Responsible Agents

The concept of meaningful human control was proposed as a crucial precondition for maintaining humans "in control, and thus morally responsible" for algorithmic decision-making [85]. Santoni de Sio and Van den Hoven [85] identified two necessary conditions for algorithms to remain under meaningful human control: *tracking* and *tracing*. The tracking condition demands decision-making algorithms to track a human agent's moral reasons relevant to the decision. The tracing condition, on the other hand, requires that someone comprehend the algorithm's capabilities as well as any of their real-world consequences. We observe a conflict between the latter condition of tracing and XAI systems.

If algorithms are explainable in a post-hoc manner, it is expected that patients understand their decisions. Explanations are designed to make algorithms explainable to a specific audience [3]. By design, post-hoc explainable AI aims to make an algorithm more understandable to patients, i.e., those interacting with them in a post-hoc manner. These explanations may thus shift perceived control from designers to those who are subjected to algorithmic decision-making.

Some of the methods presented by XAI researchers aim to empower the patients of algorithmic decision-making. For instance, Wachter et al. [104] proposed counterfactual explanations as a form of putting the recipients of decisions made by algorithms "under control." They argued that explanations—and more specifically counterfactual explanations—help patients understand why particular decisions were made, empower them to contest decisions with which they disagree, and provide courses of action they can take to achieve a different result in the future. It is worth noting that all of these goals are geared towards giving patients a better grasp of and control over an algorithm's decision-making capabilities. These approaches seem to portray patients as satisfying the tracing condition of meaningful human control. However, is this sense of control meaningful?

Rudin [83] criticizes post-hoc explanations, claiming that current approaches could create a false sense of understanding among those subjected to algorithmic decision-making. Rudin explains how saliency maps, which highlight the parts of an image that played a major role in a classification task, do not adequately describe how the model processes those parts identified as salient. Instead, this type of explanation just provides a false sense of understanding. Another example of how explanations could create a misleading sense of control is by giving explanations that are non-actionable [52]. Algorithms that tell patients they were not granted a loan because they do not have enough collateral, for instance, do not provide them with useful information.[3] Although post-hoc explanations appear to empower patients by giving them control over decision-making algorithms, this empowerment is not meaningful.

Empirical research has also shown that patients misuse AI explanations, even when they have an AI-related background [32, 53]. Algorithmic explanations can potentially deceive individuals who receive them, both intentionally and unintentionally [33], tricking them into doing things without having their interests in mind. In conclusion, this sense of control and understanding is insufficient for grounding patients as responsible for algorithmic decision-making.

We have suggested that algorithms and patients will be perceived as responsible for algorithmic decisions, but it is unclear whether this view has real-world implications. Legal systems could choose to disregard the public opinion by holding developers accountable regardless of whom laypeople choose to blame or consider morally responsible for algorithmic harm, thus bridging the responsibility

[3] Some research covers how actionable explanations could be created [51, 96, 101]; however, this field of inquiry still faces many challenges [102], such as how to evaluate explanations with patients or account for hidden features.

gap. However, doing so might hurt the adoption of possibly beneficial AI systems [11]. Detaching legal responsibility from its folk conception might create a "law in the books," which is unfamiliar to the people whose behavior legal systems aim to regulate [14]. According to empirical studies, people's reactions to algorithmic harm conflict with current legal systems [59] and adopting liability models that clash with laypeople's views may hurt the adoption of autonomous vehicles [62].

Sætra [84] has suggested that the responsibility issues posed by AI systems have inherent tradeoffs that should be assessed openly through political deliberation. Scholars have noted how laypeople's risk perceptions impacted past regulation of emerging technologies, such as genetically modified organisms (GMOs), and it is expected that laypeople will influence how AI is regulated [19]. Policymakers should be aware of the possible backlash caused by policies that go against public expectations [5]. In conclusion, people's perceptions of who should be held responsible for algorithmic harm have real-world implications for the development, deployment, and regulation of decision-making algorithms.

3.3 Designers as Responsible Agents

All the conflicts discussed above do not seem to pose a problem if designers take responsibility for the decisions made by AI systems. Designers could take responsibility for whatever harm these systems may cause, regardless of whether they are explainable. Such an approach was proposed by Champagne and Tonkens in the context of automated warfare [21]. Even if human actors do not meet the necessary conditions for backward-looking attributions of responsibility, they can take proactive responsibility for future consequences.

van Wynsberghe [99] discussed a similar approach in the context of "responsible robotics." While not directly related to algorithmic decision-making, their research emphasizes the importance of a framework that carefully considers the roles of all stakeholders throughout the development and deployment of autonomous agents. More crucially, their approach emphasizes the role of human agents in designing and deploying these systems, rather than considering algorithms (or robots) responsible. If algorithms are held responsible, they may serve only as liability shields [16].

In a similar vein, Johnson [49] argued that responsibility gaps will not arise from the technological aspects of AI systems but can only exist as a result of deliberate decisions. Such gaps will only emerge if humans developing and deploying AI systems fail to create systems that circumvent the responsibility gap, making the development of AI systems that designers can control and understand imperative.

3.4 Algorithms and Patients as Scapegoats

We agree that designers should take responsibility for the decisions made by their algorithms and be held to account if something goes wrong. First, designers often have the most assets, placing them in the best position to compensate anyone who has been harmed by these systems and thus accomplishing the primary purpose of holding entities responsible under civil law [78]. Considering that designers will profit from the deployment of AI systems, they could ensure that those harmed are compensated [20]. Second, if designers are not held responsible, they may continue developing unsafe systems in pursuit of increased profits. Being able to hold someone responsible is critical for promoting cooperation by deterring selfish behavior [37]. However, XAI systems could prove to be a tool for designers who willingly choose to shift perceived responsibility away from themselves. As previously stated, laypeople's perceptions of who is responsible have real ramifications for policymaking and the adoption of XAI systems.

Designers are unlikely to take responsibility for algorithmic decision-making to the extent that is necessary. Rubel et al. [82] highlighted how "using algorithms to make decisions can allow [...] persons to distance themselves from morally suspect actions" by shifting the responsibility for a decision to the algorithm through agency laundering, as the authors call it. The development of explainable decision-making algorithms creates a series of tools for designers to obscure their involvement and shift responsibility to the system itself. Given that explainable algorithms may be viewed as morally responsible, designers could intentionally emphasize the algorithm's role in the decision-making process. XAI systems could thus become apparent rational and intentional responsibility scapegoats. It is worth noting that explainable decision-making algorithms can be seen as morally responsible regardless of their designers' willingness to participate in agency laundering. Nonetheless, this conflict between accountability and explainability can be aggravated if designers choose to take advantage of this misleading perception.

We have argued that some types of explanations, e.g., counterfactuals, could shift perceived control of a decision-making algorithm from its designers to its patients. Following the same argument presented above concerning agency laundering, designers could highlight the role of patients in the decision-making process to obscure their own responsibility. Design methods that apparently empower patients through explanations can shift perceived responsibility for harmful outcomes to those receiving the decisions and explanations.

Explainability gives designers an exceptional degree of power by allowing them to choose what kind of information is delivered as explanations to patients [8]. As a result, designers can prioritize their interests over the well-being of individuals on the other side of the decision-making process. The concept of "dark patterns" in user interface design [13] exemplifies this power relation, in which designers can intentionally deceive users without their best interests in mind. These dark patterns can also be extended to XAI [33], shifting responsibility to patients.

Explainability could also be used as a form of ethics washing. Explainable systems could be implemented as a superficial ethical measure to avoid necessary regulation [38]. Given the numerous legal requirements for explainability in decision-making [10], designers could develop explainable systems to promote self-regulatory efforts while obscuring the need for strict regulation. The current efforts on AI ethics have been largely ineffective and vulnerable to industry manipulation [79]. Using XAI as a form of soft regulation may fail to encourage the responsible deployment of decision-making algorithms. We come back to the issue of self-regulation in Section 4.2, in which we defend hard regulation for AI systems.

It is worth noting that we do not argue that designers should *always* be held solely responsible for the harms caused by decision-making algorithms. Other actors may also be responsible depending on the circumstances. Instead, we demonstrated how designers' power over XAI systems enables them to create the impression that algorithms and patients are to blame for harmful outcomes, impacting how policy decisions are made in the real world. As the famous saying goes: "with great power there must also come—great responsibility." The threshold for shifting responsibility away from designers may need to be higher to ensure they do not escape deserved responsibility. The conflict presented in this paper can be exemplified by the burden of proof that lies with the victims of disparate treatment under US labor law. Designing AI systems that apparently empower patients to prove discrimination through post-hoc explanations shifts responsibility to individuals who do not control or understand algorithmic decision-making.

Although we have focused our discussion on the responsibility of patients and designers, another possible responsible actor is the user of decision-making algorithms. For instance, users could compensate those harmed through the profits they derive from employing AI systems. Interestingly, users remain in a position of both power over patients and subjugation to designers. While those who use decision-making algorithms can employ some of the tactics above to shift responsibility towards patients, they may also absorb responsibility from designers due to their proximity to possibly harmful algorithmic decisions [34]. Future work could explore the responsibility of users vis-à-vis patients and designers. Nevertheless, the difficulty of delimiting the responsibility of distinct actors when machines cause harm calls for proactive approaches that remove any ambiguity, as we discuss below.

4 IMPLICATIONS

The concerns raised above are mostly backward-looking, focusing on who bears responsibility for the negative consequences of algorithmic decision-making. Responsibility, on the other hand, is not necessarily backward-looking. It can also be forward-looking, emphasizing that individuals should act proactively and responsibly to the best of their abilities to ensure that future outcomes are positive [98]. Failure to attend to one's forward-looking responsibilities could lead to the attribution of backward-looking responsibility [97]. For instance, if an agent takes the responsibility (in a forward-looking manner) for ensuring that a decision-making algorithm does not discriminate against women, future audits that find this system to favor men make holding the agent responsible in a backward-looking manner appropriate (e.g., by blaming or punishing them).

Researchers have argued that concerns about responsibility gaps can be reduced or even eliminated if designers take responsibility for all stages of the development and deployment of decision-making algorithms [21, 50, 99]. However, they overlook the possibility that explainable systems could become a tool that exacerbates these difficulties, particularly (but not exclusively) if designers use them to avoid backward-looking attributions of responsibility.

Building upon previous literature proposing accountability frameworks, we discuss how these approaches could be used to mitigate the potential negative consequences of XAI systems. We build on previous suggestions by recommending concrete steps to ensure that responsibility is not wrongly shifted to patients and algorithms.

4.1 Accountability for Algorithmic Decision-Making

While accountability is frequently cited as a backward-looking notion of responsibility [98] and as a reaction to blameworthiness judgments [89], we now turn our attention to the literature defining accountability as a forward-looking notion with potential backward-looking consequences. One such example is Boven's work [12], which defines accountability as the relationship between an actor and a forum under which the actor is obligated to explain and justify their behavior. The forum poses questions and judges the actor by imposing (backward-looking) consequences, such as punishment. This definition emphasizes the actor's obligation to explain its conduct upon request, implying that actors have a forward-looking responsibility to respond to the forum. Nevertheless, if the forum deems it necessary, this requirement may lead to the imposition of backward-looking responsibilities.

Several principles have been proposed to ensure this accountability relationship can be implemented. One of these principles is traceability, which Kroll [56] advocates as a critical component in ensuring accountability for algorithmic systems. Traceability is defined as the requirements that should be satisfied so that an algorithm's outputs can be "understood through the process by which [it] was designed and developed." Kroll proposes a set of conditions (e.g., transparency in the design process, reproducibility) and tools (e.g., structured logs) to enable accountability. This approach, however, does not propose a structured framework that designers can readily and explicitly use in the development process.

We build upon Cobbe et al.'s recent proposal of reviewability as a framework for ensuring accountability in algorithmic decision-making [22]. Because algorithmic decision-making incorporates several human and temporal components, the authors argue that it should not be viewed solely as a technology. Instead, their proposal aims to understand algorithms making decisions as a broad socio-technical process. Human designers, users, patients, and other stakeholders are involved in this process not just during development and deployment but also during conception, investigation, and all other intermediate steps. The reviewability framework maintains accountability by recording "contextually appropriate information" throughout the entire socio-technical process of algorithmic decision-making. Below, we explain each step of the framework in-depth and discuss how it could handle the conflict between accountability and explainability put forth in this paper.

4.1.1 *Commissioning.*

The first step of the reviewability framework—called commissioning—addresses anything relevant prior to developing the decision-making algorithm. Cobbe et al. [22] argues that this step should define the problem algorithmic decision-making aims to solve and how it will impact society and individuals. In other words, commissioning refers to all initial human decisions that influence how the model will be developed and used. We argue that one additional and crucial question should be asked: is explainability necessary for the decision-making algorithm under commissioning?

Robbins [81] questioned whether all algorithms, particularly those deployed in low-stakes scenarios, should be explainable and argued against it. We suggest that this inquiry should also examine whether the problem under consideration requires explainability in light of the responsibility issues discussed above.

We also propose that interpretable systems should be prioritized over those that give post-hoc explanations. As discussed above, post-hoc explanations can lead to further problems concerning perceived responsibility and control. In contrast, interpretable systems place a greater emphasis on the roles of designers in developing decision-making algorithms because these systems become explainable not only after deployment but also during the development process. If post-hoc explanations are necessary for a specific problem, designers should consider which types of explanations should be provided. Explanations should not highlight the algorithm's agentic role in the process, dealing with people's perception that AI systems should be held responsible. Furthermore, explainable algorithms should genuinely empower patients rather than instilling a false sense of confidence and control.

4.1.2 Model Building. The second step of the reviewability framework focuses on the technical components of algorithmic decision-making. Cobbe et al. [22] claims that maintaining accountability requires knowledge about how data is collected and pre-processed, how the model is trained and tested, and how and by whom relevant decisions are made. Considerations about XAI and responsibility are also important throughout the model building process. Explainable systems may require different datasets and pre-processing. Although some types of explanations do not require specific datasets or annotations (e.g., feature importance explanations can be generated without additional data), others (e.g., natural language explanations) may require specific data for training. Datasets should be collected and pre-processed in ways that do not conflict with designers' forward-looking responsibilities.

Model testing is one of the most important steps in the reviewability framework to avoid any conflict between XAI and responsibility. Explainable systems should be tested rigorously with those who will use and be subjected to them once deployed. The benefits are numerous: understanding the stakeholders' view (i.e., how they perceive explainable algorithms) assists designers to prevent AI systems from being regarded as rational and intentional, which could lead to the belief that they should be held responsible; studying how users and patients interact with these systems can prevent overreliance (as done in [17]) and a false sense of control over algorithmic decision-making; testing and improving explanations can ensure they are compatible with legal and social requirements. To ensure that responsibility for algorithmic decision-making is distributed fairly, users and patients must be included in the validation of explainable systems.

4.1.3 Deployment & Investigation. The final two steps of the reviewability framework encompass the operation of decision-making algorithms (e.g., how systems are deployed, supported, and their consequences) and any subsequent investigative activity (e.g., internal and external audits). Here, users and patients play a significant role since they are intimately involved in any potential outcomes. As previously stated, these actors may also be held responsible for algorithmic decisions. Designers, for instance, cannot stop users from using algorithms for nefarious purposes and should not be held liable if users use them for unforeseeable wrongful actions. Patients may exploit explanations and abuse the system, which may be immoral or illegal.

Nevertheless, we highlight that explainable systems require particular attention concerning how responsibility should be distributed in both forward- and backward-looking manners. The deployment of XAI should not delegate responsibilities that should have been considered in previous steps of the framework to users and patients. These actors do not have complete control over decision-making algorithms and their outputs, even if explainability aims to empower them. Responsibility for algorithmic decision-making should be assigned proactively during the initial steps of the accountability framework. When the framework's final steps come into play, the respective forward- and backward-looking responsibilities of designers, users, and patients should be explicit and obvious. Such a clear division of responsibility should not be self-regulated without legally binding forces; instead, it should be codified into law.

4.2 The Necessity for Hard Regulation

We have advocated for assigning forward-looking responsibilities to designers throughout this work to avoid the conflict between explainable and accountable algorithmic decision-making. By emphasizing the prospective aspects of responsibility, designers can be held accountable for the harmful outcomes of decision-making algorithms, mitigating incorrect backward-looking attributions of responsibility to patients and algorithms. This approach underlines the non-technical components of algorithmic design, particularly during the commissioning and testing steps of the accountability framework presented above.

We have extensively discussed the literature questioning whether designers would take responsibility for their systems and argued that explainability might be used to escape responsibility. This would be comparable to what Floridi called "ethics washing" [38], in which actors engage in allegedly ethical behavior to evade regulation. What is needed to ensure that designers are held accountable for algorithmic decisions? We contend that it is hard regulation.

Jobin et al. [48] found explainability to be the most mentioned principle in a review of AI ethics guidelines and principles put forward by industry actors, policymakers, and academics. There appears to be an overreliance on ethics to ensure that responsibility is not corrupted; however, principles alone cannot guarantee ethical AI [71]. Ethics is not a substitute for hard regulation as it can be easily exploited by powerful actors and cannot ensure that principles are followed [79]. Self-regulation and soft law appear to face similar problems.

Government regulation proposals currently do not address the conflict between accountability and explainability. The European Union (EU) AI Act, the most recent attempt to regulate AI systems, does not directly address any of the responsibility issues raised by autonomous machines. Although EU expert groups have proposed potential liability models for AI systems [35] and the EU Commission has debated revising product liability laws [36], the AI Act appears to allow designers to avoid liability rather than addressing how and when they will be held liable. The current proposal allows designers to "wash their hands" by adhering to local standards through self-assessment protocols. Most crucially, these local standards are susceptible to lobbying by private organizations. By complying with private-sector standards and meeting regulatory transparency requirements, designers could shift perceived responsibility for negative outcomes to patients, who lack the necessary understanding and control over algorithms. This approach disregards the overwhelming power that designers have over XAI systems in comparison to other actors.

If responsibility frameworks are codified into law, designers will not be able to shift responsibility to other stakeholders or algorithms. Elucidating which roles and obligations designers, users, and patients should have could help mitigate the problem of the former falsely empowering others to escape blame, punishment, and other forms of backward-looking responsibility. Such a legal framework could even hold actors jointly liable if they fail to meet their obligations [103]. While it is beyond the scope of this paper to argue for a specific legal framework, we highlight the need for more research on the subject in the future. Explainable decision-making algorithms, for instance, may be required to declare their lack of agency, intentionality, and rationality alongside their explanations so that people are not influenced to hold them accountable. Such an approach would be similar to the existing proposals that mandate designers to disclose bots [100]. In conclusion, regulation can ensure that explainability and accountability coexist in algorithmic decision-making.

4.3 Blameworthy Algorithms?

People ascribe blame, responsibility, and punishment to algorithms upon harm (e.g. [41, 59, 60, 66, 67]). In this paper, we argued that explainable systems may be attributed even higher levels of backward-looking responsibilities when they provide post-hoc explanations. This effect raises the question of whether an algorithm can be held responsible for its actions. Although most academics agree that responsibility (such as blame and punishment) may be inappropriate for algorithmic systems (e.g., see [16, 26, 92]), others approach this question through a different lens [23, 44, 93]. Given the empirical evidence that humans may attribute different notions of responsibility to these systems [60]—even though they are aware that doing so is unfeasible [59]—we suspect that holding algorithms responsible may take two different routes.

One path dismisses any prospect of holding these systems responsible, arguing that the responsibility for algorithmic decision-making should be left to humans and not algorithms [99]; doing otherwise could lead to vast social and legal unrest [91]. The narrative that AI systems should be held responsible may dilute the much-needed responsibility of designers [16]. To progress along this path, AI systems should be built to refute incorrect attributions of agency, intentionality, or responsibility to algorithms. As mentioned above, explainable systems could highlight their lack of agency when providing explanations. Algorithms could be designed to oppose any attribution of mind, which has been shown to influence people's behavior towards them [29, 58].

A more controversial approach would be to invest in holding algorithms responsible, although not to the same extent or in the same way that humans are held accountable. Coeckelbergh [23] defended holding such systems virtually morally responsible to the extent that they appear to be moral agents. Similarly, Stahl [93] proposed the concept of quasi-responsibility, under which algorithms could be held responsible for fulfilling social goals; according to the author, doing so could facilitate future attributions of responsibility to human stakeholders. AI systems could also be held legally responsible if they are granted legal personhood [95]. These proposals raise the question of whether such a path is feasible and helpful. Is there any benefit in holding algorithmic systems accountable in general, or is it solely detrimental to moral, legal, and social consistency?

5 CONCLUDING REMARKS

This paper shared a concerning viewpoint that the current call for over-emphasizing explainability in algorithmic decision-making may conflict with accountability. Based on philosophical interpretations of moral responsibility and empirical research, we suggested that explainable algorithms could be seen as blameworthy and responsible. We also showed how XAI systems could shift perceived control over algorithms away from designers and towards patients and argued that this shift is mistaken. Expanding on past work questioning designers' willingness to take responsibility, we showed how algorithms and patients could become moral scapegoats that might absorb the responsibility of designers.

To avoid the potential conflict between explainable and accountable algorithmic decision-making, we have argued for a greater emphasis on designers' forward-looking responsibilities. Existing accountability frameworks should be modified to include explainability considerations at every step of AI development and deployment. XAI is an important part of the responsible deployment of algorithmic decision-making, but it should not be viewed as a panacea to all problems. The XAI field is critical for improving algorithmic decision-making, and society should be aware of how those in power may abuse it.

ACKNOWLEDGMENTS

This work was supported by the Institute for Basic Science (IBS-R029-C2) and the National Research Foundation of Korea (NRF-2017R1E1A1A01076400).

REFERENCES

[1] Heike , Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society* 6, 1 (2019), 2053951719860542.

[2] Julia Angwin, Madeleine Varner, and Ariana Tobin. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Peter M Asaro. 2016. The Liability Problem for Autonomous Artificial Agents.. In *AAAI Spring Symposia*. 190–194.

[5] Edmond Awad, Sohan Dsouza, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. Crowdsourcing moral machines. *Commun. ACM* 63, 3 (2020), 48–55.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[8] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[9] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.

[10] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2020. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* (2020), 1–21.

[11] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. *The moral psychology of AI and the ethical opt-out problem.* Oxford University Press, Oxford, UK.

[12] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework. *European law journal* 13, 4 (2007), 447–468.

[13] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. 2015. Dark Patterns-User Interfaces Designed to Trick People.

[14] Bartosz Brożek and Bartosz Janik. 2019. Can artificial intelligences be moral agents? *New Ideas in Psychology* 54 (2019), 101–106.

[15] Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8 (2010), 63–74.

[16] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.

[17] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[18] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[19] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and perceptions of AI and why they matter. (2018).

[20] Paulius Čerka, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. Liability for damages caused by artificial intelligence. *Computer Law & Security Review* 31, 3 (2015), 376–389.

[21] Marc Champagne and Ryan Tonkens. 2015. Bridging the responsibility gap in automated warfare. *Philosophy & Technology* 28, 1 (2015), 125–137.

[22] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

[23] Mark Coeckelbergh. 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society* 24, 2 (2009), 181–189.

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

[25] ACM US Public Policy Council. 2017. Statement on algorithmic transparency and accountability. *Commun. ACM* (2017).

[26] John Danaher. 2016. Robots, law and the retribution gap. *Ethics and Information Technology* 18, 4 (2016), 299–309.

[27] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies* 2015, 1 (2015), 92–112.

[28] Paul B De Laat. 2018. Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability? *Philosophy & technology* 31, 4 (2018), 525–541.

[29] Celso de Melo, Jonathan Gratch, and Peter Carnevale. 2014. The importance of cognition and affect for artificially intelligent decision makers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[30] Filippo Santoni de Sio and Giulio Mecacci. 2021. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology* (2021), 1–28.

[31] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[32] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

[33] Upol Ehsan and Mark O Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[34] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society* (2019).

[35] European Commission. 2019. Liability for artificial intelligence and other emerging digital technologies. https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en/format-PDF

[36] European Commission. 2021. Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Empty: Fostering a European approach to Artificial Intelligence. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:205:FIN

[37] Ernst Fehr and Simon Gächter. 2002. Altruistic punishment in humans. *Nature* 415, 6868 (2002), 137–140.

[38] Luciano Floridi. 2019. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology* 32, 2 (2019), 185–193.

[39] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.

[40] Matija Franklin, Edmond Awad, and David Lagnado. 2021. Blaming automated vehicles in difficult situations. *Iscience* 24, 4 (2021), 102252.

[41] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2019. Attributing blame to robots: I. The influence of robot autonomy. *Human factors* (2019), 0018720819880641.

[42] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[43] John-Stewart Gordon. 2020. Artificial moral and legal personhood. *AI & Society* (2020), 1–15.

[44] David J Gunkel. 2017. Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology* (2017), 1–14.

[45] F Allan Hanson. 2009. Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and information technology* 11, 1 (2009), 91–99.

[46] Kenneth Einar Himma. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11, 1 (2009), 19–29.

[47] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

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

[49] Deborah G Johnson. 2006. Computer systems: Moral entities but not moral agents. *Ethics and information technology* 8, 4 (2006), 195–204.

[50] Deborah G Johnson. 2015. Technology with no human responsibility? *Journal of Business Ethics* 127, 4 (2015), 707–715.

[51] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).

[52] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 228–236.

[53] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[54] Lauren Kirchner. 2020. Can Algorithms Violate Fair Housing Laws? The Markup. https://themarkup.org/locked-out/2020/09/24/fair-housing-laws-algorithms-tenant-screenings.

[55] Kirsten Korosec. 2015. Volvo CEo: we will Accept All Liability when our Cars Are in Autonomous Mode. http://fortune.com/2015/10/07/volvo-liability-self-driving-cars/.

[56] Joshua A Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 758–771.

[57] Markus Langer, 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 (2021), 103473.

[58] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *Proceedings of the 19th acm international conference on intelligent virtual agents*. 38–45.

[59] Gabriel Lima, Meeyoung Cha, Chihyung Jeon, and Kyung Sin Park. 2021. The Conflict Between People's Urge to Punish AI and Legal Systems. *Frontiers in Robotics and AI* 8 (2021), 339. https://doi.org/10.3389/frobt.2021.756242

[60] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[61] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[62] Peng Liu, Manqing Du, and Tingting Li. 2021. Psychological consequences of legal responsibility misattribution associated with automated vehicles. *Ethics and information technology* (2021), 1–14.

[63] Bertram Malle. 2006. Intentionality, morality, and their relationship in human judgment. *Journal of cognition and culture* 6, 1-2 (2006), 87–112.

[64] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[65] Bertram F Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of experimental social psychology* 33, 2 (1997), 101–121.

[66] Bertram F Malle, Stuti Thapa Magar, and Matthias Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robotics and well-being*. Springer, 111–133.

[67] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 117–124.

[68] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Frontiers in psychology* 10 (2019), 450.

[69] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[70] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[71] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[72] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[73] Satya Nadella. 2016. The Partnership of the Future. https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html.

[74] Sven Nyholm. 2018. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics* 24, 4 (2018), 1201–1219.

[75] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[76] Frank Pasquale. 2015. *The black box society*. Harvard University Press.

[77] Jairo Perez-Osorio and Agnieszka Wykowska. 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology* 33, 3 (2020), 369–395.

[78] William Lloyd Prosser et al. 1941. *Handbook of the Law of Torts*. Vol. 4. West Publishing.

[79] Anaïs Rességuier and Rowena Rodrigues. 2020. AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society* 7, 2 (2020), 2053951720942541.

[80] Marco Tulio Ribeiro, Sameer 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*. 1135–1144.

[81] Scott Robbins. 2019. A misdirected principle with a catch: explicability for AI. *Minds and Machines* 29, 4 (2019), 495–514.

[82] Alan Rubel, Adam Pham, and Clinton Castro. 2019. Agency Laundering and Algorithmic Decision Systems. In *International Conference on Information*. Springer, 590–598.

[83] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[84] Henrik Skaug Sætra. 2021. Confounding complexity of machine action: a hobbesian account of machine responsibility. *International Journal of Technoethics (IJT)* 12, 1 (2021), 87–100.

[85] Filippo Santoni de Sio and Jeroen Van den Hoven. 2018. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI* 5 (2018), 15.

[86] Thomas M Scanlon. 2009. *Moral dimensions*. Harvard University Press.

[87] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.

[88] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[89] David Shoemaker. 2011. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121, 3 (2011), 602–632.

[90] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354–359.

[91] Sheikh M Solaiman. 2017. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law* 25, 2 (2017), 155–179.

[92] Robert Sparrow. 2007. Killer robots. *Journal of applied philosophy* 24, 1 (2007), 62–77.

[93] Bernd Carsten Stahl. 2006. Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology* 8, 4 (2006), 205–213.

[94] Steve Torrance. 2008. Ethics and consciousness in artificial agents. *AI & Society* 22, 4 (2008), 495–521.

[95] Jacob Turner. 2018. *Robot rules: Regulating artificial intelligence*. Springer.

[96] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.

[97] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral Responsibility*. Springer, 37–52.

[98] Ibo Van de Poel. 2015. Moral responsibility. In *Moral responsibility and the problem of many hands*. Routledge, 24–61.

[99] Aimee van Wynsberghe. 2021. Responsible Robotics and Responsibility Attribution. *Robotics, AI, and Humanity: Science, Ethics, and Policy* (2021), 239.

[100] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.

[101] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[102] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[103] David C Vladeck. 2014. Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev.* 89 (2014), 117.

[104] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[105] Julie Weed. 2021. Résumé-Writing Tips to Help You Get Past the A.I. Gatekeepers. New York Times. https://www.nytimes.com/2021/03/19/business/resume-filter-articial-intelligence.html.