




# From Responsibility to Reason-Giving Explainable Artificial Intelligence

Kevin Baum<sup>1,2</sup> · Susanne Mantel<sup>1</sup> · Eva Schmidt<sup>3</sup>  · Timo Speith<sup>1,2</sup>

Received: 2 July 2021 / Accepted: 20 November 2021 / Published online: 19 February 2022  
© The Author(s) 2022

## Abstract

We argue that explainable artificial intelligence (XAI), specifically reason-giving XAI, often constitutes the most suitable way of ensuring that someone can properly be held responsible for decisions that are based on the outputs of artificial intelligent (AI) systems. We first show that, to close moral responsibility gaps (Matthias 2004), often a human in the loop is needed who is directly responsible for particular AI-supported decisions. Second, we appeal to the epistemic condition on moral responsibility to argue that, in order to be responsible for her decision, the human in the loop has to have an explanation available of the system's recommendation. Reason explanations are especially well-suited to this end, and we examine whether—and how—it might be possible to make such explanations fit with AI systems. We support our claims by focusing on a case of disagreement between human in the loop and AI system.

**Keywords** Explainable artificial intelligence · Reasons · Reason explanations · Moral responsibility · Responsibility gap · Decision support systems

---

This article is part of the Topical Collection on *AI and Responsibility*.

All authors share first authorship equally.

---

✉ Eva Schmidt  
eva.schmidt@tu-dortmund.de

Kevin Baum  
kevin.baum@uni-saarland.de

Susanne Mantel  
susanne.mantel@uni-saarland.de

Timo Speith  
timo.speith@uni-saarland.de

<sup>1</sup> Institute of Philosophy, Saarland University, Saarbrücken, Germany

<sup>2</sup> Department of Computer Science, Saarland University, Saarbrücken, Germany

<sup>3</sup> Department of Philosophy and Political Science, Emil-Figge-Str. 50, 44227 TU Dortmund, Germany

## 1 Introduction

Sophisticated artificially intelligent (AI) systems are spreading to evermore sensitive areas of human life. More generally, (less sophisticated) software systems, including decision support systems (DSS), which have been used for decades at this point, influence our lives in countless ways.<sup>1</sup> They are used in autonomous vehicles (Levinson et al., 2011), to support hiring decisions (Langer et al., 2018), to interpret medical images in search of indications of cancer (Kourou et al., 2015), to determine recidivism scores for convicts and help determine their sentences (Hartmann & Wenzelburger, 2021), and so forth. Many of these applications are quite advanced and err less often than humans (McKinney et al., 2020). Their use not only saves their users' time but often also helps to achieve appropriate outcomes and to prevent unwelcome or harmful consequences, e.g., car accidents or wrong medical treatments, even though these systems are not immune to error themselves.

However, many such systems are black boxes: while users can often access the systems' inputs and outputs, they cannot access or understand, let alone reenact, what happens inside the system. One reason for this is that artificial neural networks and other non-linear machine learning systems usually employ models that involve subsymbolic representations such that even developers or data science experts cannot comprehend their inner workings (Bathae, 2018).<sup>2</sup> This is often taken to be problematic especially in sensitive situations and has several bad consequences for their use: It is difficult to detect errors in the system's operations, and (arguably) neither users nor affected parties can reasonably trust such systems or make well-founded decisions based on a decision support system's recommendation, seeing as they cannot understand what underlies it. Moreover, one may worry that AI systems infringe on users' autonomy (e.g., due to nudging or outright forms of manipulation) if the systems' behavior is not interpretable to their users. And, given that it is impossible to recognize erroneous decisions or misleading recommendations, it may be difficult, and in some cases impossible, to appropriately attribute responsibility and hold anyone accountable, although in many sensitive situations it is *desirable* to be able to hold someone accountable (a claim we aim to support in Sect. 2). This problem of responsibility—the inability or difficulty of holding someone accountable even when doing so is desirable, which will be spelled out in more detail below—is the topic of this paper.<sup>3</sup>

---

<sup>1</sup> For ease of exposition, we will use “AI system” broadly to refer to software systems generally in the opening sections. Our example cases will focus on decision support systems.

<sup>2</sup> Another reason for this is of practical nature. Often such systems are proprietary and companies are afraid to lose their competitive advantage if they offer insights into the systems' inner workings. As we are interested in the general problem that comes with black boxes, we focus on principled black boxes and leave such practical considerations for policy and regulatory experts.

<sup>3</sup> A prominent position addressing problems of responsibility in this area is the account of meaningful human control developed by Santoni de Sio and collaborators (Santoni de Sio & Van den Hoven, 2018; Mecacci & Santoni de Sio 2020). We are less optimistic than they are about the prospects of ensuring responsibility in the case of fully autonomous systems, and so will argue for the need to keep a human in the loop (in many cases). Further, explainability of AI systems will be a central component of our strategy to deal with responsibility gaps, whereas Santoni de Sio et al. propose that their conditions of tracking and tracing can be met by way of strategies that do without explainability. Another contrast is that they focus on the control condition of moral responsibility, whereas our approach will draw on the epistemic condition. Finally, their account of meaningful human control builds up specifically on Fischer and Ravizza's (1998) notion of guidance control; we take our account to be more easily compatible with a broader range of approaches to moral responsibility understood as a kind of accountability. That said, their account is congenial to our proposal in many ways. Keeping a human in the loop and providing them with reason explanations—as we will propose—may well be *one* way of ensuring meaningful human control.

That the opacity of AI systems gives rise to these problems is intuitively plausible. Arguably, solutions have to put users and people affected by automated or algorithmically supported decisions in a position to understand what underlies the decisions of the systems. In other words, explanations must be provided.<sup>4</sup> The goal of research in explainable AI (XAI), consequently, is to open the black box, or at least to make it more translucent and perspicuous (Langer, Oster, et al., 2021). XAI, understood in a broad sense, is pursued by researchers from a range of disciplines.

This multidisciplinary nature comes with a lot of different perspectives and focuses. For instance, a whole host of papers revolve around problems like those mentioned in the previous paragraphs; they provide arguments for XAI from the broader context of morality or society in general (e.g., Asaro, 2015; Binns et al., 2018; Cave et al., 2018; Floridi et al., 2018; Langer, Oster, et al., 2021; Lipton, 2018; Wachter et al., 2017). However, these discussions do not always tell us how *exactly* we can get from a need for reasonable trust, human autonomy, accountability, responsibility, or the like, to a requirement for explainable AI systems. Moreover, they typically do not tell us which kinds of explanations should be given to meet these concerns.

At the same time, there is a broad variety of technically minded papers from computer science introducing and discussing concrete methods for coaxing explanations out of AI systems (e.g., Bach et al., 2015; Kim et al., 2018; Montavon et al., 2017; Ribeiro et al., 2016; Selvaraju et al., 2017). These papers, however, simply presuppose that their results will help to fulfill the proclaimed requirement. This is not surprising, since they are usually not informed regarding the richness of the nature of concepts like *explanations*, *explaining*, *interpretation*, or *understanding* (Miller, 2019; Miller et al., 2017).

Finally, there are a few papers, such as Wachter et al. (2018), Zerilli et al. (2018), and Miller (2019), that strive to provide a more philosophically and psychologically informed picture of the explanations that AI systems should give. However, despite proposing particular kinds of explanations (viz., intentional, counterfactual, or contrastive explanations), they remain silent on whether, or how, explanations of these kinds meet the needs which motivate the call for explainable AI to begin with, such as enabling reasonable trust, human autonomy, or responsibility. To sum up, there is little discussion of whether and how specific forms of explanations—to be provided by technical tools mentioned in the previous paragraph—deliver precisely what the arguments from a societal perspective in favor of XAI demand.

Against this backdrop, our aim is to combine ideas from all three types of papers: We begin by defending and clarifying the claim that there is a desideratum to be able to hold an individual morally responsible for morally problematic AI-supported decisions or actions in Sect. 2. We then argue that such decisions should often be made by a human in the loop who receives recommendations from a decision support system (Sect. 3). Next, by appealing to the epistemic condition on moral responsibility, we substantiate the claim that the outputs of many such decision support

---

<sup>4</sup> For a general framework for relating the above and other societal desiderata of various stakeholders to explainability, see Langer et al. (2021a). Specifically, for the case of (reasonable) trust and trustworthiness, see Kästner et al. (2021). Furthermore, see Chazette et al. (2021) for a general model of the impact of explainability on various social and technical phenomena.

systems must be explainable for the human in the loop for her to bear responsibility (Sect. 4). By appealing to cases of disagreement between DSS and human in the loop, we argue that explanations of a certain kind—viz., reason explanations—are especially suitable for enabling morally responsible decision-making (Sect. 5). We conclude with some practical challenges for developing reason-giving XAI systems (Sect. 6).

## 2 The Challenge of Adequate Responsibility Attribution

The call for XAI is often motivated by appeal to worries about high-stakes situations<sup>5</sup> in which moral harms may result from opaque systems, among them the worry that missing explainability leads to an inability to hold anyone accountable, or responsible, if something goes wrong. Let us first turn to why exactly it is important to be able to (appropriately) ascribe responsibility when AI systems are operating, and then turn to the question what is needed to be able to do so. For this, we need to understand what is meant by “responsibility” in the relevant contexts, starting with clarifying the concept of responsibility. To do so, we compare and contrast it with the related legal concept of accountability.

Problems of legal accountability are central to the legal concerns with XAI, for instance, in connection with discussions of an alleged EU Right to Explanation (Wachter et al., 2017).<sup>6</sup> Unfortunately, the term “accountability” is used in a variety of ways in this debate.<sup>7</sup> Decision-makers (and agents generally) are accountable, in the sense in which we are interested, when they can appropriately be *held* to account, i.e., when it is appropriate to demand that they explain or justify their conduct or, further, when they deserve reprimand or punishment, given that their decisions or actions are unlawful (Zarsky, 2013; Edwards & Veale, 2017; see Duff, 2007 and 2019 for a nuanced picture of criminal responsibility).

This legal term is structurally quite similar to philosophical notions of moral responsibility (Talbert, 2019).<sup>8</sup> Moral responsibility for an action, as discussed in philosophy, is often spelled out in terms of the agent’s blame—or praiseworthiness for the action, where this is understood in terms of its being fitting to have certain emotions towards the agent such as resentment, indignation, anger,

<sup>5</sup> As a current example, we refer to the proposal of the Artificial Intelligence Act of the European Commission: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>. Accessed September 29, 2021.

<sup>6</sup> See the General Data Protection Regulation (EU) 2016/679 (GDPR), <https://data.europa.eu/eli/reg/2016/679/oj>. Accessed September 29, 2021.

<sup>7</sup> For instance, “accountable” is sometimes used as indicating that persons that are accountable have to be ready to justify their decisions or actions upon request; as more or less synonymous with “explainable” (Kroll et al., 2017); or as concerned with fair and effective governance (de Laat 2018; Perel & Elkin-Koren 2016,). We will put these uses to one side here.

<sup>8</sup> One author who insists on structural similarities between moral and criminal responsibility is Duff (2007; 2019). Duff (2007) discusses the widely held claim that criminal responsibility presupposes moral responsibility. For opposition to this claim, see Shoemaker (2012; 2013).

or gratitude (Strawson, 1962). This approach has been developed in contemporary debates in various forms (see, e.g., Wallace, 1994; Watson, 1996; McKenna, 2012; Pereboom, 2014; Shoemaker, 2015). The corresponding notion of responsibility is often called “accountability,” and is distinguished from other notions of moral responsibility such as attributability or answerability. Though much of what we argue may hold for different forms of responsibility, we are concerned with responsibility primarily in the sense of appropriate praise—or blameworthiness, as exemplified by Shoemaker’s (2015, 113) notion of accountability: “One is an accountable agent just in case one is liable for being a fitting target of a subset of responsibility responses to one – a subset organized around the paradigm sentimental syndrome pair of agential anger/gratitude – in virtue of one’s quality of regard.”<sup>9</sup> In the following, when speaking of responsibility, accountability is what we have in mind.

Moreover, while it may be that moral responsibility presupposes causal—or more broadly—counterfactual responsibility, it goes beyond that concept: It may be that a moral harm would not have come about had I acted differently, but still I am not blameworthy, for example, because I was not aware of what I was doing. Relatedly, computer scientists will say that certain components of a system are accountable or responsible for its failure, i.e., the failure is counterfactually dependent on the performance of these components, but this does not amount to a claim that they are morally responsible or accountable in the sense we are concerned with (Chockler & Halpern, 2004; Halpern & Pearl, 2005).

Here is an everyday example for the kind of moral responsibility we are interested in. Imagine that human resources (HR) manager Herbert, who is tasked with deciding which applicant will get an important management position in the company, disqualifies April, a black female applicant, because of her race and gender. Herbert is not seriously psychologically impaired. It is therefore appropriate to respond to Herbert’s action by blaming or reproaching him for his behavior, and in this case even by taking up legal measures against him for discriminating against the applicant (see, e.g., Title VII of the US Civil Rights act of 1964 or the German *Allgemeine Gleichbehandlungsgesetz*). That is to say, Herbert is morally responsible and legally accountable for his action.

The use of AI systems can challenge the ascription of this kind of responsibility. Suppose Herbert’s company employs a fully automated hiring system to screen, rank, and

<sup>9</sup> As the inclusion of *praiseworthiness* (for one’s good deeds) indicates, moral responsibility is not an exact analogue of *legal* accountability, which is associated with one’s bad deeds only.

We assume that moral responsibility can be ascribed for decisions as well as actions, but also for their outcomes, and that in the cases we discuss, what is at issue is responsibility both for actions and the decisions that underlie these actions, and also their outcomes. Consequently, we speak of moral responsibility for actions and moral responsibility for decisions interchangeably in this paper. Where appropriate, we also speak of responsibility for the outcomes of an action or decision.

Further, agents are morally responsible for actions *under a description*. In the case we focus on in this paper, it may be that Herbert the HR manager is responsible for rejecting April’s application, but not responsible for discriminating against her—in terms of the “action under a description” terminology, he may be responsible for his decision or action under the description “rejecting April’s application,” but not under the description “discriminating against April.” We say more about this below.

select job applicants. Assume that the system ranks April in the last place and excludes her from the further hiring process. Now maybe this ranking was decisively influenced by the fact that April is a Black woman, or some other irrelevant information. If this is the case, this intuitively raises multiple concerns. One is the question of unfair algorithms and algorithmic bias and discrimination (e.g., Garcia, 2016).<sup>10</sup> Another is the worry that no one can be held morally responsible or legally accountable for excluding April, for there was not *anyone* who excluded her. Matthias (2004) calls this a “responsibility gap”.<sup>11</sup> This responsibility gap, understood as an accountability gap, will be the focus of our paper.<sup>12</sup> We will concentrate on cases of responsibility for biased AI-supported decisions since there is much discussion of algorithmic bias.<sup>13</sup>

Let us sketch two motivations for closing the moral responsibility gap, for making sure that there is a person who can be properly held responsible for such morally problematic decisions. We do so by focusing on the case of Herbert and April. On the one hand, there is a motivation from incentives: If someone like Herbert is morally responsible for the problematic decision or action, this means that he can fittingly be blamed for it. It is then, at least *pro tanto*, just to express blame or even to establish legal sanctions (McKenna, 2012, though there may be exceptions). This will plausibly motivate him to be diligent in making up his mind about whether to (follow the system’s recommendation and) disqualify the black female applicant to avoid negative consequences for himself. Such an incentive for diligent decision-making may lead to better hiring decisions and less wrongdoing (for empirical evidence, see Fehr & Gächter, 2002).<sup>14</sup>

<sup>10</sup> We cannot enter into the debate about discriminatory and unfair models here. Roughly speaking, a model is unfairly biased against members of a certain group if (and only if) it treats its members quantitatively unequally (to their disadvantage and without justification), for instance, by making certain classificatory errors more or less frequently. This particular unequal treatment may either have immediate causes in input data, for example, when the system directly refers to April’s characteristic of being Black, or it may be caused indirectly, by giving certain proxy variables undue weight, such as April’s Alma Mater, her hobbies, or her zip code. A range of technical issues may be behind such biases in systems, e.g., when sensors perform worse for darker skin colors. Furthermore, there may be problems with the quality of training data, for example, when the data does not adequately represent the population or when it is a result of a biased social process. Some unfair treatment is generally mathematically unavoidable (Chouldechova 2017; Lepri et al., 2018).

<sup>11</sup> For overviews, see, e.g., Johnson (2015) or Noorman (2020).

<sup>12</sup> The terminology in the responsibility and responsibility gap debates is not unified. For instance, our accountability gap corresponds largely to Santoni de Sio and Mecacci’s (2021) culpability gap, but not to their accountability gaps.

<sup>13</sup> Related worries about moral responsibility can be raised by other cases: For instance, one might worry about how properly to allocate responsibility in case a system simply does not work reliably. Furthermore, there may be a DSS that recommends a decision that is optimal given only its available information. However, the human decision-maker has further information available that, put together with the system’s information, shows that this decision is terrible. Yet without access to the information that the system relied on, the decision-maker is unable to put two and two together, so to speak, and therefore unable to figure out what would be the best decision overall in the situation. It seems wrong to hold the decision-maker responsible for making a bad decision in this example. We discuss a case of this kind in Sect. 5.

<sup>14</sup> While (possibly unjust) expression of blame and punishment could of course also incentivize agents who cannot be held responsible, such incentives do not seem palatable to many, and going this far is not necessary as long as someone can be held accountable to whom incentives can be more appropriately applied. Furthermore, an agent who meets a central condition of responsibility is better placed to comply with incentives to avoid harming others (as we argue in Sect. 4).

We acknowledge that this argument needs further details in order to evade counterarguments. For instance, it has been put to us that one may always be able to find someone responsible for producing or employing the DSS if it discriminates against applicants, and that person will have an incentive to be diligent that no discrimination arises. However, as we point out in the next section, if the system's discrimination is not *foreseeable* to anyone, there may be no one bearing indirect responsibility of this kind. Furthermore, it might be that only someone at the company who developed the system can be held indirectly responsible but nobody at the companies that employ the system. Then, the system might be applied carelessly by a great number of users who need not bother as long as, for instance, the system is not taken off the market. In this case, there would be no incentive for applicants to avoid wrongdoing in hiring decisions.

On the other hand, there is a motivation of justice: If April suspects—or finds out—that she was discriminated against because of her race and gender, it would intuitively be desirable to enable her to blame someone for wronging her. It would be desirable to make it possible for her to get justice, in the sense of a *person responsible* for discriminating against her owning up to the fact that they did something wrong. She should be able to be fittingly angry with someone and to express this anger by demanding of a responsible decision-maker that they acknowledge their wrongdoing, that they apologize, make amends; it would be desirable to make it possible that they get sanctioned. To motivate this further, imagine that the responsibility gap cannot be closed. Then April's situation is morally equivalent to the situation of another agent, call her Berta, who has been harmed by a natural disaster: Both April and Berta are harmed, nobody is responsible, and nobody is blameworthy. However, April's and Berta's situations are intuitively different. Many people were involved in setting up and using the system that harms April, but no human is involved in harming Berta. And it seems that this makes a difference in terms of justice, for Berta really cannot justly blame anyone, but intuitively April *should* be able to appropriately blame someone and may reasonably desire to do so.

Of course, here too one may raise doubts, for instance, by questioning whether justice requires being able to angrily blame someone or just being able to do something in the vicinity. Maybe all that is needed is someone who is *answerable* in the sense explicated in Shoemaker (2015, 82),<sup>15</sup> i.e., someone who is able to cite their reasons for the action and who is thereby liable for being a fitting target of responses like agential regret or pride in virtue of their quality of judgment. In our concrete case, this person would be expected to admit and regret a discriminatory hiring decision. Such answerability would not imply accountability.

Arguably, that there is someone who is answerable might already be helpful to some degree to ensure justice for the wronged applicant. However, in the case of discrimination and other offensive treatment, it would further be desirable for an agent like April to be able to fully hold someone accountable. While it seems

---

<sup>15</sup> Further explications of answerability that do not imply accountability can be found in Smith (2005) and Scanlon (2008, chap. 4). Duff (2007) takes answerability to consist, roughly, in someone's being an appropriate target of the request for an explanation. We thank an anonymous referee for pressing us to address the counterexamples to this and the following arguments.

right that April deserves an explanation, she should also be able to be fittingly resentful for being disadvantaged based on her race and gender, and to be able to call for moral sanctions in terms of blame. This indicates that accountability and not just answerability is relevant (see Shoemaker, 2011, 616 and 621).<sup>16</sup> We acknowledge that these questions can be debated further. However, since our main focus is an argument to the effect that explanations are often the best way of closing the moral responsibility gap, it is sufficient for our purposes here to present these initial motivations that could be spelled out further. In our view, the argument from incentives and the justice-based argument provide a compelling rationale for a desideratum to avoid high-stakes situations in which no one can be held responsible.

### 3 Why We Need Someone in the Loop

But how to make sure that there is a person who can properly be held responsible? In this section, we will argue that, if we want to ensure that a human can bear responsibility for morally problematic decisions, we often cannot—and, in fact, should not—delegate these to fully automated systems. Instead, we should keep a human in the loop: AI systems should be used merely to supply recommendations about what to do, but the final decision should be left to a human decision-maker—in our example, to Herbert.<sup>17</sup> However, keeping a human in the loop is, as we argue in the next section, not sufficient to ensure that there is someone who can bear responsibility. But before we can turn to the question of what is missing for a sufficient condition—and how this relates to explainability of a certain kind—we want first to give a convincing argument for requiring a human in the loop.

The obvious alternative to keeping a human in the loop to bear responsibility would be to find someone else at the company, someone who decided to purchase the system or (one of) the developers of the system and to allocate moral responsibility for the specific fully automated decision to that person. What speaks against this alternative? In a fully automated decision process, no one made the decision or was able to influence it directly. So no person can bear *direct* responsibility for (the outcome of) the fully automated decision. A person at the company or a developer could, at most, bear responsibility *indirectly* given that the decision was fully automated. Indirect responsibility can be ascribed to an agent for an outcome where she is directly responsible for something else—such as

<sup>16</sup> The same argument could be phrased in a more victim-centered way: Plausibly, in a situation in which applicants are systematically discriminated against, many of them will *desire* to be able to at least hold decision-makers or their companies accountable for discriminatory decisions, to be able to resent and blame them. This is a good reason to ensure that someone can indeed be held responsible.

<sup>17</sup> Human in the loop cases contrast with cases where a human is on the loop: Here, a human supervisor is informed about the decisions of an AI system before they are put into effect and can step in if need be. We will focus on the “in the loop” scenario for sake of simplicity. What we say for this scenario can be modified and applied to many, though not all “on the loop” scenarios.



her own ignorance or loss of control—which led to that outcome. In such cases, this “something else” is her fault, for instance, because she did not do enough to meet obligations to stay informed or in control (Mele, 2021; Rosen, 2003; Zimmerman, 1997).<sup>18</sup> One might think that someone will bear indirect responsibility for the fully automated decision by being responsible directly for employing—or designing—a faulty system, so we can also blame them, indirectly, for the particular decision made by the system.

The proposed assignment of indirect responsibility, however, runs up against an especially nasty variant of the problem of many hands (Thompson, 1980; van de Poel et al., 2015). The problem of many hands, as we understand it here, is quite generally that, in a complex situation, in which the contributions of many agents lead to moral harm, such as when large corporations and companies cause a problem, it is difficult to allocate direct or indirect moral responsibility to anyone in particular. The problem of many hands has an epistemic and a metaphysical dimension: On the one hand, it is concerned with difficulties in determining who is morally responsible and, on the other hand, with difficulties with respect to whether anyone actually is responsible.<sup>19</sup> It further has a practical-political dimension: Even if there is someone who is—directly or indirectly—responsible, complex situations with many contributors lend themselves to obfuscation, making it easy for companies and other agents to let themselves off the hook.

Adding a fully automated AI system to the mix compounds the problem. Suppose that a level 5 self-driving vehicle<sup>20</sup> kills a pedestrian. Should the blame and thus the moral responsibility for the accident be allocated to the company who built the car, to the company who supplied the car’s LiDAR (light detection and ranging), or to the company who owned and employed the vehicle for the mission it was undertaking, etc.? If one of the companies is held responsible, which person at the company is to bear responsibility? It appears that an already complex situation here is made even more confusing by the involvement of an autonomous AI system (Awad et al., 2018; Coeckelbergh, 2020; de Laat, 2018; Mittelstadt et al., 2016; Nissenbaum, 1996; Sparrow, 2007). Having a human in (or at least on) the loop, by contrast, alleviates the problem of many hands by providing at least one easily detectable and plausible candidate for bearing the direct responsibility for a particular decision that caused harm. After all, if the car were operated by a human who could reject any recommendation or decision of the system, that person is an

<sup>18</sup> A related notion is that of “tracing,” which was introduced by Fischer and Ravizza (1998, 49).

<sup>19</sup> Problems of many hands can arise in virtue of independent decisions of several different agents, which are each sufficient for a harmful result. Since the result is overdetermined, no one agent could have prevented it by acting differently, so it seems each agent is off the hook. Such problems can also arise when each contribution of several agents by itself would have been harmless, but all actions in combination lead to a morally problematic result. Here, it may be that no agent by herself could have prevented the result, so again it is difficult to ascribe moral responsibility to anyone.

<sup>20</sup> See SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, [https://www.sae.org/standards/content/j3016\\_202104/](https://www.sae.org/standards/content/j3016_202104/). Accessed June 23, 2021.

obvious candidate for blame.<sup>21</sup> This is not to deny that it may be important to allocate additional indirect responsibility to the companies involved if these could and should have done more to prevent accidents.<sup>22</sup>

One central way in which the problem of many hands may arise is that, because of the many agents involved in a situation, agents contributing to a harm are unable to foresee that their combined actions will lead to a problematic outcome. We will explore the role of knowledge for responsibility in more detail in the next section, but even pre-theoretically, it seems problematic to hold someone responsible for a harmful outcome when it was not foreseeable by them, and so when they were not at fault for not foreseeing it. Transferring this worry to the case of fully autonomous AI systems like the hiring system used by Herbert's company, it may well be the case that all people who might bear moral responsibility indirectly for the output of an AI system blamelessly lack relevant (fore-)knowledge regarding the system's output. This may be true, e.g., of the developers of an AI system, of people at an accreditation agency, and of the companies employing the system. Even a thoroughly tested and generally reliable system may give a problematic output when certain features of the situation to which it responds combine in an unusual way (Edwards & Veale, 2017).<sup>23</sup> Focusing on the issue of bias, systems cannot easily be tested before or during employment with respect to bias because the biases for which they would be tested concern protected classes, such as sexual orientation, which often is not available in the data to the developers (Lepri et al., 2018). Furthermore, bias might be hidden in the statistics—even though a system, say, puts women at a disadvantage as compared to men, the overall statistics regarding its performance may look unproblematic. Thus, there are likely many cases in which neither developers nor the persons employing the system can be expected to foresee particular harmful future outputs of a system. And so they cannot properly be held (indirectly) responsible for them.

A final problem for allocating indirect responsibility is that there may be cases in which, although we may succeed in finding a person (e.g., developers or employers of the AI system) to whom we can ascribe indirect responsibility for a particular output, it would be wrong to hold them responsible nonetheless. For it could be that the system is the best system that could have been developed—e.g., it is the least biased hiring system that it is possible to design—but still there are some fringe cases in which its output is biased, e.g., in that it puts Black queer women from a low socioeconomic background at a disadvantage. To put it differently, it may

<sup>21</sup> In the Uber car accident in Tempe, Arizona, in March 2018, a self-driving vehicle owned by Uber was operated by a human on the loop who was supposed to intervene in case of emergency; otherwise, the car was in self-drive mode. Here, the operator is the most salient candidate for bearing responsibility, although not the only one. See Will Pavia (March 21, 2018), Driverless Uber car “not to blame” for woman's death, *The Times*, <https://www.thetimes.co.uk/article/driverless-uber-car-not-to-blame-for-woman-s-death-klkbt7vf0>. Accessed June 23, 2021.

<sup>22</sup> A human in the loop is ideally positioned to realize that a DSS is not working properly, e.g., by exhibiting bias in many recommendations. She is thereby able, in principle, to recognize that the developers of a system are at fault for creating a malfunctioning system, for which they can be held morally responsible.

<sup>23</sup> This concern relates in interesting ways to the issue of intersectionality discussed in feminist theory: The experience of discrimination made, e.g., by a black woman, is different and potentially worse than that of a white woman or a black man, since different categories/dimensions of discrimination combine (Cooper 2016).

be that the overall great performance can only be achieved at the price of allowing some suboptimal outputs in rare cases. We can even imagine that the system is much better overall than a human decision-maker would be. If so, it seems that the developers or employers of the system have done nothing wrong, and so cannot be blamed. Nonetheless, it would be desirable to have someone who can bear responsibility for the individual biased hiring decision and its morally problematic outcome.

Even if a candidate for indirect responsibility could be identified (contrary to the epistemic dimension of the problem of many hands), the unpredictability of problematic outputs and the issue of overall optimal performance may prevent that candidate from bearing responsibility. This holds both for the developer of the AI system and for a customer who relies on an accredited system that she leaves to operate by itself. Bearing moral responsibility for a particular output of an AI system, even indirectly, requires that there is someone who is able to foresee it and who cannot evade blame because they did the best they could. As argued, these conditions will often not be met.

With this argument in place, we need to add a qualification. Fully automated or autonomous AI systems may be acceptable in some cases. For instance, two of the most pressing and widely debated applications discussed in the context of responsibility gaps are autonomous driving and (lethal) autonomous weapons. Autonomous driving typically involves no human at all or at most a human on the loop. Similarly, while the mode of operation for drones in general has been moving more and more from human in the loop to human on the loop setups,<sup>24</sup> lethal autonomous drones involve at best a human on the loop, who can interfere with the decisions of some autonomous system that identifies potential targets. In both cases, the time available to make a decision may not be sufficient for an effective handover, let alone an “explained handover,” even if such a handover is technically possible. In light of this, a human in the loop and concurrent explainability of an output to this human may not be all things considered the best way to go, even if this entails *pro tanto* undesirable responsibility gaps. For example, assume that autonomous vehicles<sup>25</sup> prove to be clearly superior to human drivers in certain contexts, so that critical situations only occur in a fraction of cases, while the time for an (explained) handover is too short. If this is the case, it may be that a fully automated set-up is in some cases significantly better than one involving a human in the loop, so that it *may* be all things considered permissible to leave corresponding responsibility gaps open.<sup>26</sup> However, we believe that this is true only of a limited number of cases (e.g., some cases with extreme time pressure or very low stakes), so that our argument gets a grip in a significant number of cases.

---

<sup>24</sup> See United States Air Force Unmanned Aircraft Systems Flight Plan (May 18, 2009): [https://irp.fas.org/program/collect/uas\\_2009.pdf](https://irp.fas.org/program/collect/uas_2009.pdf). Accessed September 29, 2021.

<sup>25</sup> Let us make explicit that we make the following argument for autonomous vehicles, but *not* for lethal autonomous weapons systems. Here, we believe that the stakes are too high for the advantages of a fully automated setup to outweigh the disadvantages of not being able to hold anyone directly responsible.

<sup>26</sup> Similarly, responsibility gaps may be acceptable in low-stakes decision situations, e.g., involving movie recommendation services on streaming platforms. While it may be *prima facie* desirable to have a responsible agent for every single recommendation (as they may be discriminatory as well), plausibly the stakes are simply too low to justify the effort.

Next, are we not letting developers and companies employing decision support systems off the hook too easily? This is not so. Note the following two features of our argument. First, the strength of our claim: Our aim is to establish that keeping a human in the loop (and providing them with explanations) is *one* good way of ensuring that we can properly hold someone responsible. We suggest that, in some contexts, this may be the best or even the only way to go, but leave open that there may be other ways for ensuring responsibility more suitable for other situations (and some of these ways may rely on explainability or other perspicuity enhancing capabilities *after* the fact, to determine what went wrong in the relevant situation, see Sterz et al., 2021).

Second, the scope of our claim: Our focus is on moral responsibility, and how to ensure that there is an agent who can properly bear it in the context of AI-supported decision-making. Whether the same argument applies to related phenomena such as legal accountability is a further issue beyond the scope of this paper. One suggestion is that, even if it is not possible properly to ascribe moral responsibility to the developers or employers of an AI system, we may still be able to hold them accountable by law (e.g., by imposing a strict liability for damages arising from the operation of a car on its registered keeper). The current debate over the German law for regulating automated driving, which has been criticized for making vehicle owners liable for damages instead of manufacturers, indicates that similar problems arise in legal contexts.<sup>27</sup> Finally, we allow that there may be cases in which it is justifiable not to enable moral responsibility, e.g., where affected parties are compensated for not having someone to hold responsible.

Overall, we conclude that the allocation of indirect responsibility is often infeasible. Instead, we then need a person who is presented with the output during use and has the chance to interfere—a human in the loop. Since a human in the loop is made knowledgeable of the recommendation during use and makes the relevant decision herself, she is a candidate for direct responsibility for the outcome.<sup>28</sup>

#### 4 Connecting Responsibility to Explainability

A human in the loop is a candidate for responsibility, but there are further requirements to properly allocate responsibility to them. This is where the demand for explainability comes in. As Floridi et al. put it, ensuring “that the technology – or, more accurately, the people and organizations developing and deploying it – are held accountable in the event of a negative outcome, ... would require ... some *understanding* of why this outcome arose” (2018, p. 700, our italics).<sup>29</sup> To have such understanding, the human in the loop, at the time of the decision, needs access to

<sup>27</sup> Gerald Traufetter (May 7, 2021), Vorstoß von Verbraucherschützern – Hersteller sollen bei Unfällen mit autonomen Fahrzeugen haften, *Der Spiegel*, <https://www.spiegel.de/auto/autonomes-fahren-hersteller-sollen-bei-unfaellen-haften-fordern-verbraucherschuetzer-a-78df0b3a-0002-0001-0000-000177426963>. Accessed June 22, 2021.

<sup>28</sup> In the following, when speaking of responsibility, we refer to direct moral responsibility, if not stated otherwise.

<sup>29</sup> For a useful discussion of this issue in the medical context, see Robert David Hart (September 10, 2018), “Who’s to blame when a machine botches your surgery?” <https://qz.com/1367206/whos-to-blame-when-a-machine-botches-your-surgery/>. Accessed June 23, 2021.

an explanation of the DSS's recommendation and possibly its overall functioning. Our aim in this section and the following is to motivate and substantiate the claim that explainability is needed to make the human-in-the-loop solution work, and to investigate what kind of explainability would do the job well. We do so by focusing on what a human in the loop needs in order to meet a standard condition for moral responsibility: This is, at bottom, an explanation of the system's output.

The foundation for our reasoning lies in a necessary condition on direct moral responsibility which is widely discussed in the philosophical debate—the *epistemic condition* (Noorman, 2020; Rudy-Hiller, 2018). According to it, an agent is not directly morally responsible for an action unless she is aware, or in a position to be aware, of what she is doing, of the (probable) consequences of her action, of its moral significance, or of alternative options available to her. For instance, an agent who flips the switch to turn on the light and who thereby electrocutes her neighbor by an unfortunate combination of circumstances that was not foreseeable is not directly responsible for the harm caused.

One way to make this distinction more tangible is to resort to a coarse-grained view of actions according to which one action can be picked out under a range of different descriptions (Anscombe, 1962; Davidson, 1963). In the example, the agent's action can be described as flicking the switch *or* as turning on the light *or* as electrocuting the neighbor. Since she is not in a position to be aware that her action is one of electrocuting her neighbor, she is not directly morally responsible for it under that description, though she may still be responsible for flicking the switch. For Herbert and April, the crucial question then, in the context of the epistemic condition, is not whether Herbert is responsible for rejecting April, but whether he is responsible for discriminating against April.<sup>30</sup> In light of this distinction, the epistemic condition can then be spelled out thus:

(Epistemic Condition) An agent is morally responsible for her action or decision only if she has sufficient epistemic access to it. *That she has sufficient epistemic access to it entails at least that she is in a position to know the action under relevant descriptions.*<sup>31</sup>

The epistemic condition on moral responsibility can be used to provide two motivations for making decision support systems explainable—the first motivation will be introduced by appealing to an initial case, and the second by appealing to a fleshed-out version of this case. Our *initial* case is the hiring case in which HR manager Herbert is a human in the loop and makes the final hiring decision, but does not have an explanation of the hiring system's recommendation. Assume that, before his company started to employ the decision support system, Herbert used to be the HR manager who competently and responsibly made hiring decisions for his company,

<sup>30</sup> We will leave out the qualification 'under a certain description' in the following, except where it is necessary for our argument. We assume that the relevant description is clear from context. Plausibly, the agent's role (such as being an HR manager) helps to determine the descriptions under which we want to hold her responsible for her action or decision.

<sup>31</sup> Note that the question of what descriptions exactly are relevant goes beyond the scope of this paper. This determination might well be a function of the specific context of the decision-situation. For an idea of how to operationalize context-sensitive societal desiderata more generally, see (Köhl et al., 2019; Langer et al., 2021a).

and that he will continue to do so, using the DSS's output as one source of support. We focus on human in the loop cases like Herbert's, in which the decision-maker relies on a DSS and no other AI systems play a role.

Imagine that Herbert decides to exclude April's application because the hiring system recommended doing so. Imagine further that the system's recommendation is due to its bias against Black female applicants, but that, since it is an accredited system, Herbert justifiably believes that it has no such problems. Herbert is therefore not indirectly responsible for discriminating against April—he is not to blame for being unaware of the system's bias. If he is responsible, he must be directly responsible, which requires his being in a position to know what it is that he is doing, its probable consequences, and its moral significance. As described, if he does not have access to what moved the DSS to provide its recommendation, then his AI-supported decision will be made without him being in a position to know these things. Herbert is aware that he rejects April's application, and so he is aware of his action under that description. But he is not in a position to know that what he is doing, under another description, is to discriminate against her. Nor is he in a position to know that he unfairly rejects her application and that this is an act of moral wrongdoing. Consequently, he is not morally responsible for discriminating against April.

Once a meaningful explanation of the recommendation is available to the decision-maker, we can more easily bridge the responsibility gap. For instance, assuming that the system discriminates against April immediately based on her race and gender, then, if Herbert has access to this fact, he does have access to—is in a position to know—the fact that to reject her application on this basis is to discriminate against her; and that it is unfair and an act of moral wrongdoing. But even in the case where the system discriminates against April based on a learned correlation involving some otherwise innocent proxy variables such as, say, April's Alma Mater, her hobbies, and her zip code, explanations may enable Herbert to get the right kind of epistemic access. For the proxies will typically be either suspicious or seemingly irrelevant. In both cases, Herbert should doubt the system's recommendation: If the system indicates that it considers the combination of April's Alma Mater, her hobbies, and her current zip code to be particularly crucial, this may catch Herbert's attention: Is this not one of the historically Black colleges and universities? And is that not a primarily Black neighborhood?

In any case, an explanation allows Herbert to become suspicious and to pay particular attention to the role played by other factors. Herbert can then check, if necessary, whether candidates with otherwise similar profiles are rated similarly. In this case of proxy-based discrimination, Herbert may not be sure that discrimination is present, but given sufficient background knowledge and awareness of the danger of discrimination by models, he can develop an initial distrust and at least begin to consider that other descriptions of the situation might be relevant. He is therefore in a position to know at least that a decision that follows the system's recommendation *may very well be* discriminatory. So, while explanations may not guarantee in all cases that the epistemic condition on moral responsibility is met, they clearly facilitate its fulfillment.<sup>32</sup>

<sup>32</sup> If the required background knowledge of the danger of discrimination by models is unavailable to Herbert, he may fail to be responsible for discriminating against April, since the proxies may then not raise his suspicions. But then maybe someone else is (indirectly) responsible for not having ensured that Herbert has what it takes to fulfill his role competently.

Let us turn to our *second* motivation. At least on one way of fleshing out Herbert's situation further, his epistemic situation is even worse than has become apparent so far. Our fleshed-out scenario shows that, if a decision-maker cannot tell why a DSS provided the recommendation it did, then there may be situations, particularly situations of disagreement between system and decision-maker, in which he cannot tell whether his decisions bring him closer to his goals. As a consequence, he is unable to guide his decisions so as to pursue these goals, or to execute his intentions in acting. This gives rise to an especially threatening way in which an agent lacks epistemic access to his action, and thereby also lacks moral responsibility for it.

Here is the fleshed-out scenario. Imagine that Herbert, at the end of a lengthy selection procedure, is presented with a list of three applicants that the DSS ranks as the top candidates; the system recommends keeping them in the running for the position. April did not make the list, but made it into the top ten. However, Herbert, by going through the top ten applications independently, counted her among the top three applicants beforehand. So we have a case of disagreement between the system's recommendation and Herbert's initial judgment. Since there is no explanation of the system's recommendation available, Herbert cannot reasonably resolve the disagreement.

Here is how this might happen: Say that his own assessment of April's qualities is due to good, but not conclusive reasons—she has more relevant work experience than most; received great grades in her studies at Yale; speaks a foreign language, which is useful but not absolutely necessary for the job; and has work experience abroad. (By saying that his reasons are not conclusive, we mean that they are weak enough that he may reasonably question his own judgment if the system gives a contrary recommendation.) On the other hand, the system was accredited to be reliable by a trustworthy watchdog organization, though Herbert is aware that systems of this kind may have hidden bugs or biases. In this situation, the system's countervailing recommendation leaves open both the possibility that Herbert correctly assesses the situation and the system is mistaken *and* the possibility that the system has a superior understanding of the situation, and Herbert is in the wrong. In the first possibility, the system's recommendation may be due to some kind of bug, or to its bias against women of color; in the second possibility, the system's recommendation may be due to the fact that it has access to information Herbert does not have, or detects patterns that Herbert misses. Say that the system relies on all of Herbert's reasons for taking April to be among the top three candidates (her great grades from Yale, her foreign language competences, etc.). However, it has detected that applicants with these qualifications *taken together* tend to move on to other, better jobs very quickly. So the system detects a pattern which turns what would otherwise be great reasons for hiring a candidate into a reason against hiring her.

This illustrates that, in a particular situation, Herbert may be unable to tell whether he is in one of two relevant cases:

*Case 1* The system's recommendation is mistaken and Herbert's assessment is right.

*Case 2* The system's recommendation is correct and Herbert's assessment is wrong.

Given that the two cases are indistinguishable to him, he cannot reasonably resolve the disagreement. For he cannot compare or reconcile his own and the system's reasons for or against keeping April in the running, and so cannot figure out which reasons are superior, e.g., by weighing them against each other. Consequently, if he decides to keep her in the running, this decision is arbitrary; but if he decides

to exclude her from the short list, that decision is also arbitrary.<sup>33</sup> The lack of access to the system's reasons undermines Herbert's ability to come to a well-founded all-things-considered judgment about which applicants to keep in the running.

In light of his inability to come to a well-founded all-things-considered decision, Herbert is then unable to competently pursue his goal. Say he is genuinely trying to find the best candidate for this prestigious, responsible position at his company. Since he is unable to tell which is the proper means to doing so—keeping April in the running or excluding her—he is thereby unable to respond to pertinent reasons in pursuit of his goal. In other words, he cannot properly guide his decisions in light of his goals, so as to execute his intentions. This undermines his ability to find the best candidate or to reach various related goals. Imagine Herbert is instead trying to damage the company by hiring an unsuitable candidate. Again, since he cannot tell whether it is his or the system's assessment of April that is right, he is unable to tell whether excluding April would be a good means to pursuing this goal, and this undermines his ability to guide his hiring decision in response to pertinent reasons.

In the fleshed-out scenario, Herbert is especially epistemically impaired: He is not in a position to know either of his options under the relevant descriptions. He cannot tell whether, if he complies with the system's recommendation, his decision is one that wrongs April; but neither can he tell whether, if he goes with his own initial assessment, his decision can be described as one of harming his company. In this fleshed-out version of the scenario, then, Herbert's access to his decision is undermined in a more severe way. Because of this more wide-ranging epistemic disconnect, Herbert is not directly morally responsible for his AI-supported decision.<sup>34</sup>

<sup>33</sup> But can a human in the loop not decide non-arbitrarily by relying on the system's (or his own) past track record? There may be situations where this is possible, and for these, we concede that the problem is mitigated. However, this leaves many situations in which both the human in the loop and the DSS have equally good or bad track records, where this information does not help. Moreover, it can be hard to determine a track record: For some tasks, such as finding the best applicant out of a pool of applicants, it may be difficult to figure out whether they have been performed successfully, and so also whether a subject or a system has a good track record with respect to the task. We thank an anonymous reviewer for posing this question.

<sup>34</sup> One might use this line of thought to raise trouble for responsibility in three further ways, which we cannot develop here, but would like to at least mention: (1) One might connect it to another widely discussed necessary condition for moral responsibility, the control condition (Talbert 2019), if control is understood to require reason responsiveness. Our case of disagreement arguably undermines the decision-maker's ability to recognize a certain class of reasons, those which lead to the recommendation of the DSS. One could spell this out by appealing to a constraint included in Fischer and Ravizza's (1998, 64) account, viz., that the agent has to be able to act for the sufficient reason that favors his action. If the agent is unable to access the system's reasons, then he is unable to act for them (Mantel 2018).

(2) It might be argued that our case undermines Herbert's (backward-looking) responsibility via undermining his ability to meet his forward-looking responsibility as an HR manager. Plausibly, in virtue of his professional position, Herbert has the task—and therefore the obligation—of finding the best candidates for jobs at his company, without unfairly relying on, e.g., candidates' membership in certain groups. If he fails to meet this professional obligation, he can appropriately be held accountable for this failure. In cases of disagreement, Herbert's inability to know what decisions he has available under relevant descriptions and his consequent inability to competently pursue his goals, undermine his ability to fill his professional role, and so he lacks forward-looking responsibility. So, given that backward-looking responsibility hinges on forward-looking responsibility (Duff 2019), Herbert cannot be held responsible for his hiring decisions in these cases either.

(3) One might explore whether, on some Strawsonian picture, it could be said that, for the agent to be morally responsible, their quality of will must be *expressed* in the action in a way it cannot be expressed in Herbert's case without having access to the system's reasons.



Of course one might object that cases of disagreement are insignificant outliers. Typically, the decision-maker will agree with the system's recommendation. However, this objection renders the use of decision support systems obsolete. If the system's recommendation allows for well-founded decision-making only where it supports what the decision-maker would choose anyway, then it is pointless to combine a DSS with a human in the loop for the hiring decision. From Herbert's perspective, adding the DSS does not improve his decision-making; from the perspective of the company, keeping a human in the loop does not add an advantage over employing a fully automated system. The system can lead to better decision-making exactly by way of disagreement with the decision-maker where there is room for changing his mind. So, *exactly when it counts*—when the decision-maker has reasons that are not conclusive, and the system makes a recommendation that is potentially better than his take on the situation—the system undermines the decision-maker's epistemic access to his decision, and thus his moral responsibility.

Without explainability, we face a dilemma for human in the loop scenarios: It is *either* pointless to have the system provide a recommendation to the human decision-maker (in cases where human and system agree, or when the decision-maker has conclusive reasons anyway), *or* the lack of explainability undermines his epistemic access to his decision and thus the moral responsibility which the human in the loop is supposed to bear (in cases where human and system disagree, while the human has non-conclusive reasons). Now the second horn of the dilemma is due to the fact that the decision-maker has no access to why the DSS provided a certain recommendation. If he had a suitable explanation of the system's recommendation available, so that he would be able to compare his reasons with the system's reasons, he would be in a better position to figure out whether it is the system's or his own assessment of the situation that is correct. So, he would be able to resolve the disagreement in a non-arbitrary way, thereby be able to make the hiring decision that best suits his goal (finding the right person for the job), and thus be in a position to know his decisions and actions under the relevant descriptions. We conclude that, in many cases of disagreement where the decision-maker's reasons are non-conclusive, he is in a position to bear direct responsibility for his decision just in case he has a suitable explanation of the system's recommendation available.

To sum up, a human decision-maker needs explanations. These enable responsible AI-supported decision-making by enabling the agent to meet the epistemic condition in cases like the ones discussed in this section.

## 5 The Advantages of Reason Explanations

Which form should an explanation take to ensure that decision-makers are morally responsible for their AI-supported decisions? While different kinds of explanations could enable responsibility when properly interpreted by human decision-makers, reason explanations are particularly well-suited for this job. They are the ones that humans typically use when trying to understand and explain action, when exchanging justifications for actions and recommendations, and when trying to resolve disagreements (Alvarez, 2010; Hieronymi, 2011). Just like human experts would

provide reasons for their recommendations, so should decision support systems. In this section, we spell out how reason explanations help to resolve different kinds of disagreement between humans in the loop and DSS and what kind of reasons are needed for the job.

Before returning to the disagreement case and illuminating what reason explanations for decision support systems should look like, let us first clarify what reasons are and which kinds of reasons figure in reason explanations. In the philosophy of action, reasons are categorized by the distinction between normative and motivating reasons (Alvarez, 2017; Hieronymi, 2011; Mantel, 2018). We here apply—without defending it—this widely accepted philosophical distinction to the recommendations of decision support systems. *Normative* reasons are facts that objectively favor or disfavor an action (such as the action recommended by a DSS). All normative reasons, taken together, make the action right or wrong. For instance, the fact that eating vegetables is healthy counts in favor of my eating vegetables. Applied to decision support systems, we may say that normative reasons are the facts which favor or disfavor a DSS's recommendation and the recommended action. When a system's input data contains information that fits the facts and supports the recommended action over another, we can say that the system has available normative reasons favoring a certain recommendation.

Although ideally a DSS has normative reasons available, reason explanations should focus on *motivating* reasons instead, because systems can make mistakes. A motivating reason is a consideration that an agent relies on in acting, a consideration “for which someone does something, a reason that, in the agent's eyes, counts in favor of her acting in a certain way”—whether or not it is a fact and actually favors the action (Alvarez, 2017). Motivating reasons stand at the intersection between explanation and justification insofar as they help to explain the output in the light of what the decider took to justify or favor it (Hieronymi, 2011). Unlike normative reasons, motivating reasons can include merely apparent facts, i.e., non-obtaining states of affairs or false propositions that the agent falsely takes to obtain (Dancy, 2000; Schmidt, 2018). For instance, that spinach is a good source of iron is a merely apparent fact. Even though it is not the case that spinach is a good source of iron, this can be the reason which motivates me to eat spinach—since I mistakenly believe that spinach is a good source of iron, in my eyes, this favors the action, and it is the light in which I act. If a motivating reason is not mistaken, we say that it corresponds to a normative reason.

We suggest what one might call a functional picture of motivating reasons, on which “favoring in an agent's eyes” is not interpreted as entailing awareness. We talk of motivating reasons more loosely to pick out information which plays a certain role in determining the output of a system, e.g., in whether or not it recommends a certain action. With this functional characterization in mind, it becomes feasible to transfer reasons to decision support systems. A DSS can then be described as providing recommendations on the basis of reasons available to it, or, to put it differently, as treating something within its inputs as reasons for its recommendation. For it can be correct that the system provides a certain recommendation because it has certain information (i.e., motivating reasons) available. Note that this does not

yet commit us to the claim that there is a form of (non-deflationary) reasoning to be found within that system.

Turning next to *reason explanations*, a reason explanation explains an action in terms of an agent's motivating reasons—that is in terms of the information or misinformation that led her to the action. Ideally then, a reason explanation of a system's recommendation will include only the information on which the system relied in producing its output—the information contained in the data available to the system on which it relied in providing its recommendation. The explanation refers to the information which *actually* contributed to the system's coming to a particular recommendation, and not to confabulations. This is not to say, however, that the reason explanation refers to all the information that made a contribution to the recommendation or decision. Although agents may be aware of a huge number of *pro* and *contra* considerations and may be led to an action by such a bundle of reasons, most reason explanations of human action focus on just one or a few contextually relevant motivating reasons. Even if a DSS takes into account much more information than a human would in providing a recommendation, this complexity therefore does not rule out providing a simple reason explanation for its recommendation, for such explanations typically do not require to name all of the motivating reasons but only the most relevant ones. What it does require, of course, is singling out *some* contextually relevant pieces of information, and especially the most significant ones.

Typically, humans have no access to the reasons on which a DSS bases its recommendations or to the roles they play in producing these recommendations. In order to be able to offer reason explanations, therefore, one would ideally be in a position to examine the actual decision-making processes of the system and to present the involved reasons and inferences accordingly. But presuming this would be naïve. More and more DSS are based on modern developments in AI. Neural networks and support vector machines, which operate on high-dimensional data spaces, seem to elude precisely this form of access and understanding of the internals, which has earned them the title of “black boxes” (Bathae, 2018).

There are several obstacles to providing reason explanations for the recommendations of such DSS: First, there might simply be no decision process in the relevant sense. Perhaps a system learns to solve a particular task without any representation or structure at all. The concept of tacit knowledge (compare Polanyi's paradox, Autor, 2014; Polanyi, 1966) and the distinction between “knowing how” and “knowing that” (Bathae, 2018) may be relevant in explaining how such systems can prepare recommendations and make decisions without relying on reasoning processes. Importantly, though, such systems will still offer systematic, non-random outputs relative to inputs. Otherwise they would just be random generators. But they are not—many such systems work really well, i.e., reliably provide extremely useful and fitting outputs.

Second, however, our inability to provide explanations for a DSS's outputs may be rooted in an epistemic deficiency: We simply do not gain access to hidden reasoning processes. A typical explanation of this is that the reasons and processes are represented in a distributed manner at the subsymbolic level of artificial neurons

(Goodfellow et al., 2016). But if these processes elude our access, we can certainly not easily provide them or the reasons involved therein.

And even if we could access such reasoning processes, there is a third reason why we might fail to provide the right kind of reason explanations: It is possible that the actual reasons and reasoning processes simply cannot be processed and grasped by humans, i.e., that they are incomprehensible to us (Armstrong et al., 2012). This could be the case because they are too high-dimensional to be visualized or otherwise too complicated to be suitably represented. Alternatively, such systems might use a conceptual scheme that is too different from ours to be expressible in human terms and that therefore resists translation (for doubts concerning the meaningfulness of this last claim, see Davidson, 1973).

However, these obstacles do not render the pursuit of reason explanations an impossible, hopeless endeavor. For one, the reason explanations we give for human actions are useful even though they are often approximations of far more complex processes (and may similarly face problems such as members of different cultures or linguistic communities having different conceptual schemes, or the connectionist structure of and processing in the human brain). For another, even complex reasons and reasoning processes—given they *do exist*—can in principle be approximated. A satisfactory account of how this is possible lies beyond the scope of this paper, but discussions in the philosophy of science regarding the non-factivity of understanding (Elgin, 2007), surrogate reasoning (Contessa, 2007), as well as idealization and approximation with respect to models (Frigg & Hartmann, 2020; Potochnik, 2007; Strevens, 2017) indicate a way forward. In case of their *non-existence*, we can generate sufficiently good explanations externally by methodically interpreting the systematic behavior of the DSS.<sup>35</sup>

Indeed, many existing explainability methods do something along these lines. LIME (Ribeiro et al., 2016) is a good example of this. To explain the prediction for some input, LIME approximates a complex model locally around this input by a simpler model that can then easily be explained. In other words, what is used to explain the prediction is not the original model (that may elude understanding because of its high-dimensionality), but a simpler model (with fewer dimensions) that behaves like the original model for inputs similar to the input in question. Similarly, we could generate reason explanations for a complex system by constructing a simpler system that locally approximates (relative to some observed prediction or recommendation) the original DSS. To do so, we would have to construct the simpler system in such a way that we can properly attribute reasons to its decision-making process, while staying sufficiently faithful to the behavior of the original DSS. That is to say that the simpler system has to give more or

---

<sup>35</sup> Remember our functional understanding of reasons and reasoning. Following Boghossian's (2014) discussion of inference, we can think of AI systems as reasoning in the following deflationary sense: As long as they are disposed, given certain premises, to come to certain conclusions (in other words, as long as they have the disposition to give certain recommendation where relevant information is available to them), they have the functional profile of a reasoner, and we will thus say that they undergo reasoning processes.

less the same recommendations for sufficiently similar inputs (for a suggestion along these lines, see Baum et al., 2017).<sup>36</sup>

Reason explanations allow humans to assess a system. For example, such explanations make it in principle possible to assess whether the system's motivating reasons are, or correspond to, normative reasons that favor the recommended action.<sup>37</sup> A well-working system responds to facts which are normative reasons.<sup>38</sup> This means that the system's recommendation will be actually favored by the facts, and the system will, in general, be robustly responsive to the facts, so it is no mere luck that it provides a good recommendation, but it does so across a broad range of situations.<sup>39</sup> By contrast, if the system is completely off track, a reason explanation of its outputs may not mention any normative reasons at all but only non-obtaining or irrelevant considerations. Even so, the explanation would be very useful to the person in the loop—for instance, by revealing that the system is malfunctioning in a particular way.

Let us apply these thoughts to our example of a hiring system.<sup>40</sup> If a hiring system recommends hiring a certain candidate for a job, a normative reason for that recommendation would be any fact that indeed obtains and that objectively favors hiring the candidate, such as the fact that she is very clever, well educated, and works accurately even under pressure. Many normative reasons may not be available to a DSS, for instance, because a CV does not fully disclose a candidate's personality and capacities. But given useful and human-processable reason explanations, a human in the loop should be able to incorporate further reasons available to her into *her* reasoning process—which is a crucial part of her role.

Now return to our disagreement case. Suppose Herbert has what he thinks are good, but not conclusive reasons for keeping April in the running, whereas the system excludes her from the top three applicants. What exactly is it that Herbert needs so he can resolve this disagreement in a non-arbitrary way? He needs to be able to figure out which party to the disagreement is in the wrong, by figuring out whether one party overlooked normative reasons that the other recognized, relied on motivating reasons that were mistaken, or gave the reasons too much or too little weight,

---

<sup>36</sup> Although they are not reason explanations themselves, the explanations provided by LIME can be used as good starting points to generate or extract reason explanations. Similarly to many other explainability approaches, LIME explanations highlight the most salient features for a prediction. In an image of a giraffe, for instance, we expect its long neck to be a salient feature (compared to other animals) that will be highlighted by LIME. Now, the reason for the prediction can be extracted from what is highlighted. If the reason for the prediction “giraffe” is that the image depicts an animal with a long neck, we can judge that the model works for inputs similar to this. By contrast, if the reason for the prediction “wolf” is that snow is depicted in an image, we can judge that there is an error in the model.

<sup>37</sup> For a discussion of identity and correspondence, see Mantel (2014) and Mantel (2015).

<sup>38</sup> That a system is “well-working” can mean, e.g., that it tracks reality in a reliable way, that it is accurate, or that it is robust (in a technical sense).

<sup>39</sup> See Mantel (2018) about the absence of luck in acting for normative reasons. To appropriate Fischer and Ravizza's (1998, 69) term, the system is then “reasons-receptive.”

<sup>40</sup> As may be apparent to our readers, the following toy example uses an unrealistically oversimplified DSS to emphasize the general point. In reality, DSS make their recommendations dependent on dozens, hundreds, thousands, or even millions of parameters. The art of representing adequate, but at the same time useful reason explanations in a way appropriate to the context requires not only technical but also philosophically and logically-conceptually underpinned empirical-psychological research. We say a little more about this in the outlook. Thank you to one of our reviewers for pressing us on this point.

or the like. In many situations, there are further features which modify normative reasons by disabling, attenuating, or strengthening them. They, too, need to be considered, as we will show below. In sum, the decision-maker has to be in a position to figure out whether the following possibilities are at the root of the disagreement:

*Disagreement of fact* System and decision-maker represent reality differently. They treat different propositions as facts or assign different uncertainty measures to propositions.

*Disagreement of relation* System and decision-maker treat different purported facts as favoring (or disfavoring) a course of action, they assign different strengths to favoring (or disfavoring) relations, or they treat purported facts as interfering with favoring (or disfavoring) relations, e.g., by disabling or attenuating them.

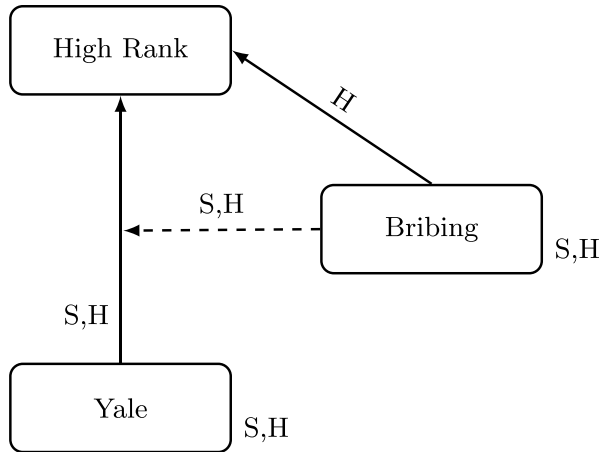
The human in the loop wants to check the motivating reasons on which the system relies and thus to identify disagreements of relation and disagreements of fact. For instance, there is a *disagreement of fact* if Herbert falsely believes that April has more relevant work experience than the others, whereas the system does not. If the system's rejection of April is explained by appeal to the reason that she has less work experience, this will enable him to double-check his information and to reasonably revise his original judgment (or to stick with it, if the mistake lies with the system's representation of the facts).

A *disagreement of relations* is in place, for example, if Herbert and the hiring system assume that the same facts obtain, but assign them different roles or different relations to the decision in question. This may be the case when they take the reasons in a situation to have different weights or to count in favor of different, mutually exclusive decisions; or if they disagree over whether these facts are reasons, or over whether some features modify the given reasons, as intensifiers, attenuators, enablers, or disablers. To illustrate a disagreement based on different assessments of modifiers, suppose that the DSS detects a pattern that Herbert misses: Applicants with April's (otherwise positive) traits taken together tend to move on to other, better jobs very quickly. Here, facts that would individually be great reasons to hire April, together constitute a reason *against* hiring her.

To take a more complex case (see Fig. 1), both Herbert ("H") and the DSS ("S") may be apprised of the fact that April was accepted at Yale after her mother bribed the school ("Bribing").<sup>41</sup> The system counts this fact as a disabler (dashed arrow): Given that she was accepted at Yale because of a bribe, the system does not take the fact that she studied at Yale ("Yale") to be a reason in favor of hiring her ("High Rank"). Moreover, it takes the fact that her mother bribed the school as a reason against hiring her, taking it as evidence of a lack of moral integrity. By contrast, Herbert, a person of low moral character, believes that the fact that April's mother is willing and able to use bribes to pave her daughter's way as a (prudential) reason to hire her (continuous arrow). To his mind, this fact indicates that April is from a rich, well-connected family and will therefore be an asset to the company. So, while he

<sup>41</sup> We are referencing the 2019 college admissions scam which was widely reported in the news, e.g., Jennifer Levitz and Melissa Korn (March 14, 2019), The Yale Dad Who Set Off the College-Admissions Scandal, *The Wall Street Journal*, <https://www.wsj.com/articles/the-yale-dad-who-set-off-the-college-admissions-scandal-11552588402>. Accessed April 14, 2021.

Fig. 1 Yale bribing example



also thinks of this fact as a disabler, he treats it not as a reason against hiring her, but as a reason that favors hiring her.

Again, if Herbert receives the system's assessment that the fact that her mother bribed the school disables the fact that April went to Yale as a reason to hire her, he is in a position to integrate this knowledge into his own decision-making. For instance, he might then discount the system's recommendation because the system is blind to the importance of coming from a well-connected family; or he might come to realize that it is more important to fill this position with someone who made it into an excellent university without bribery, and comply with the system's recommendation. Either way, he will meet the epistemic condition with respect to his decision, and bear moral responsibility for it.

Generally speaking, disagreements are typically resolved by taking into account the reasons of the other party. The decision-maker needs a grasp of what reasons the system operated with and how it treated pieces of information, e.g., as reasons or as disablers. This is to say, a reason explanation needs to state explicitly what pieces of information served as reasons for or against a certain recommendation and what pieces of information served as modifiers of reasons. Furthermore, the explanation needs to include the strengths of these reasons. If the decision-maker has access to this information, he can reassess his information about the facts as well as his treatment of them as playing different roles such as those of reasons, disablers, attenuators, etc. He can then come to an all-things-considered decision that integrates all relevant facts in a coherent way, weighing the relevant reasons against each other, and he is then in a position to know his decision under the relevant descriptions and, thus, to be morally responsible for the decision.<sup>42</sup>

Reason explanations are the explanations we typically use to communicate reasons. By contrast, other forms of explanations would seem to make resolving disagreements much harder. Imagine that Herbert is provided with the following explanation of why

<sup>42</sup> It should be noted that—as the reader may already suspect—Herbert's use of the system's reason explanations not only makes him responsible in the sense of accountable but also puts him in a position to be answerable (see Shoemaker 2011, 616) for the resulting decision. For it enables him to cite the reasons that he took to justify the decision when he made it. Our focus here, however, is on the role that explainability plays for responsibility not in the sense of answerability, but of accountability.

the system recommended against April: If her mother had not bribed the school, it would have recommended to hire April. Such a *counterfactual* explanation, as suggested by Wachter et al. (2018), indicates that the facts mentioned in the explanation were taken either as reasons against hiring April, or as disablers of other reasons to hire her (and in the example, both are the case). But as this example illustrates, he may still be unable to tell which of the two roles a fact played (reason against or disabler, or both)—with one role he agrees, with the other he does not—and it will take extra work for him to assign the facts their proper roles and to integrate them correctly into his own reasoning.

## 6 Open Questions and Future Work

In this paper, we have argued that, to close responsibility gaps, we often need a human in the loop who is in a position to bear direct responsibility for her—AI-supported—decisions. However, for a human in the loop to be in a position of directly responsible decision-making, she needs to have the right kind of epistemic access to relevant features of her action. We have argued that the epistemic condition on moral responsibility often cannot be met by the human in the loop if she has no access to the system's motivating reasons for its recommendation. We have explained how meeting the epistemic condition translates to certain abilities in practice, first and foremost to the ability to recognize and resolve disagreements of different kinds between man and machine. And we have argued that reason explanations are theoretically well-suited to restore epistemic access, supplying a background picture of motivating and normative reasons from the philosophy of action, which we started to transfer to decision support systems and their recommendations.

However, all this can only be a starting point. Several empirical and technical tasks remain on the path to useful machine-generated reason explanations. In a further empirical step, one could ask *how many* and *which* motivating reasons need to be provided (especially when a DSS processes a great amount of information) and *how they need to be presented* in the explanation of a recommendation such that the human can best use the explanation. This includes the issue of how the strength of reasons or different roles such as disabling, attenuating, or intensifying should be represented. In explaining a system's recommendations, what is needed is an explanation that users can understand and often one that they can comprehend quickly. When humans give explanations, they intuitively present information selectively and focus on the information that seems relevant in the context of a given question. Providing more information than necessary can be distracting, and it leaves the recipient of the explanation the time-consuming task of singling out the bits that are most relevant. This can be counterproductive.<sup>43</sup> Hence, the explanation presented initially would ideally involve only the most relevant motivating considerations, while flagging their respective roles. Less relevant motivating reasons would be provided only upon a request to give more detailed information. But which reasons are relevant and why? This calls for further, especially psychological and normative research.

---

<sup>43</sup> For instance, more information about a digital application process can easily undermine the organizational attractiveness (Langer et al., 2018). It seems not too far-fetched to assume similar counterproductive effects due to too many or poorly presented reasons.



Relatedly, it seems problematic to try to provide *general* principles of which reasons will be the most relevant elements of a reason explanation. For the relevance of a reason is not determined solely by its significance within the specific reasoning process, but might well be a function also of the aims or background knowledge of the human who receives the explanation. This seems to call for an interactive way of explaining that allows the human to dive deeper into the why, a typical problem for human–computer interaction.<sup>44</sup>

This brings us to the question of how, quite generally, reason explanations of AI systems' outputs should be represented to their addressees. What is a suitable data structure for reason explanations? We not only need a way to represent reasons, but also their relations and quantitative information like their weight or potentially involved uncertainty. Formal, graph-based approaches to reasons (Horty, 2012) as well as argumentation and dialectical frameworks (Amgoud & Prade, 2009; Baum et al., 2018, 2019; Dung, 1995) might lead the way. Further research along these lines calls for input from both theoretical and practical computer scientists.

In short, we believe that the endeavor of equipping decision support systems with the ability of giving reason explanations is not only imperative, but opens up several interesting and highly interdisciplinary lines of research for the future.

**Acknowledgements** We thank audiences at the Rhine-Ruhr Epistemology Meeting; at the lecture series Gratwanderung Künstliche Intelligenz (Dortmund); at the Colloquium of the Department of Values, Technology and Innovation (Delft); at the reading group for theoretical philosophy (Dortmund); at the EIS Kolloquium (Saarbrücken/Dortmund); at the research colloquium of the DCLPS (Düsseldorf); and at the reading group for philosophy and technology (Frankfurt School of Finance & Management) for helpful discussion. Special thanks go to Benjamin Kiesewetter, Sebastian Köhler, Sara Mann, Leonhard Menges, Filippo Santoni de Sio, Sarah Sterz, Christine Tiefensee, Johannes Weyer, and Simon Wimmer, for lots of great input. Finally, we are grateful for the in-depth comments of two anonymous reviewers, which helped us to greatly improve our argument.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work was partially supported by VolkswagenStiftung as part of Grant AZ 98510, 98512, and 98514 EIS—Explainable Intelligent Systems (see <https://explainable-intelligent.systems/>), and partially funded by DFG grant 389792660 as part of TRR 248, CPEC (see <https://perspicuous-computing.science>).

## Declarations

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

<sup>44</sup> Langer et al. (2021b) and Schlicker et al. (2021) provide some insights into context factors that may impact what is required of an explanation to be of use to a human in a certain context.

## References

- Alvarez, M. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199550005.001.0001>
- Alvarez, M. (2017). Reasons for Action: Justification, Motivation, Explanation. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/>
- Amgoud, L., & Prade, H. (2009). Using Arguments for Making and Explaining Decisions. *Artificial Intelligence*, 173(3–4), 413–436. <https://doi.org/10.1016/j.artint.2008.11.006>
- Anscombe, G. E. M. (1962). *Intention*. Blackwell Press.
- Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4), 299–324. <https://doi.org/10.1007/s11023-012-9282-2>
- Asaro, P. M. (2015). The Liability Problem for Autonomous Artificial Agents. In *Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposia*, 190–194. AAAI.
- Autor, D. (2014). Polanyi's Paradox and the Shape of Employment Growth. *National Bureau of Economic Research Working Paper Series*. <https://doi.org/10.3386/w20485>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7). <https://doi.org/10.1371/journal.pone.0130140>
- Bathae, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*, 31(2), 889–938. <https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathae.pdf>
- Baum, K., Köhl, M. A., & Schmidt, E. (2017). Two Challenges for CI Trustworthiness and How to Address Them. In *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*. <https://doi.org/10.18653/v1/w17-3701>
- Baum, K., Hermanns, H., & Speith, T. (2018). From Machine Ethics to Machine Explainability and Back. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*. <https://www.powver.org/publications/TechRepRep/ERC-POWVER-TechRep-2018-02.pdf>
- Baum, K., Hermanns, H., & Speith, T. (2019). Towards a Framework Combining Machine Ethics and Machine Explainability. In *Proceedings of the 3rd Workshop on formal reasoning about Causation, Responsibility, and Explanations in Science and Technology (CREST)* (pp. 34–49). <https://doi.org/10.4204/eptcs.286.4>
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3173951>
- Boghossian, P. (2014). What Is Inference? *Philosophical Studies*, 169(1), 1–18. <https://doi.org/10.1007/s11098-012-9903-x>
- Cave, S., Nyruup, R., Vold, K., & Weller, A. (2018). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, 107(3), 562–574. <https://doi.org/10.1109/jproc.2018.2865996>
- Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *IEEE 29th International Requirements Engineering Conference (RE)* (pp. 197–208). IEEE. <https://doi.org/10.1109/RE51729.2021.00025>
- Chockler, H., & Halpern, J. Y. (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22, 93–115. <https://doi.org/10.1613/jair.1391>
- Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press. <https://doi.org/10.7551/mitpress/12549.001.0001>
- Contessa, G. (2007). Scientific Representation, Interpretation, and Surrogative Reasoning. *Philosophy of Science*, 74(1), 48–68. <https://doi.org/10.1086/519478>
- Cooper, B. (2016). Intersectionality. In Disch, L., & Hawkesworth, M. (Eds.), *The Oxford Handbook of Feminist Theory* (pp. 385–406). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199328581.013.20>
- Dancy, J. (2000). Practical Reality. *Oxford University Press*. <https://doi.org/10.1093/0199253056.001.0001>

- Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, 60(23), 685–700. <https://doi.org/10.2307/2023177>
- Davidson, D. (1973). Radical Interpretation. *Dialectica*, 27(3–4), 313–328. <https://doi.org/10.1111/j.1746-8361.1973.tb00623.x>
- de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, 31(4), 525–541. <https://doi.org/10.1007/s13347-017-0293-z>
- Duff, R. A. (2007). *Answering for Crime*. Hart Publishing.
- Duff, R. A. (2019). Moral and Criminal Responsibility: Answering and Refusing to Answer. In Coates, D. J., & Tognazzini N. A. (Eds.), *Oxford Studies in Agency and Responsibility Volume 5* (pp. 165–190). Oxford University Press. <https://doi.org/10.1093/oso/9780198830238.003.0009>
- Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games. *Artificial Intelligence*, 77(2), 321–357. [https://doi.org/10.1016/0004-3702\(94\)00041-x](https://doi.org/10.1016/0004-3702(94)00041-x)
- Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a “Right to Explanation” Is Probably Not the Remedy You Are Looking for. *Duke Law & Technology Review*, 16, 18–84. <https://doi.org/10.2139/ssrn.2972855>
- Elgin, C. Z. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42. <https://doi.org/10.1007/s11098-006-9054-z>
- Fehr, E., & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature*, 415(6868), 137–140. <https://doi.org/10.1038/415137a>
- Fischer, J. M., & Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press. <https://doi.org/10.1017/CBO9780511814594>
- Floridi, L., Cowlis, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Frigg, R. & Hartmann, S. (2020). Models in Science. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>
- Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, 33(4), 111–117. <https://doi.org/10.1215/07402775-3813015>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <https://www.deeplearningbook.org>
- Halpern, J. Y., & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4), 843–887. <https://doi.org/10.1093/bjps/axi147>
- Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sciences*, 54(2), 269–287. <https://doi.org/10.1007/s11077-020-09414-y>
- Hieronymi, P. (2011). XIV-Reasons for Action. *Proceedings of the Aristotelian Society*, 111(3), 407–427. <https://doi.org/10.1111/j.1467-9264.2011.00316.x>
- Horty, J. F. (2012). Reasons as Defaults. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199744077.001.0001>
- Johnson, D. G. (2015). Technology with No Human Responsibility? *Journal of Business Ethics*, 127(4), 707–715. <https://doi.org/10.1007/s10551-014-2180-1>
- Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169–175). IEEE. <https://doi.org/10.1109/REW53955.2021.00031>
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668–2677). <https://proceedings.mlr.press/v80/kim18d/kim18d.pdf>
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. In *IEEE 27th International Requirements Engineering Conference (RE)* (pp. 363–368). IEEE <https://doi.org/10.1109/RE.2019.00046>

- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G. & Yu, Harlan (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, 165, <https://ssrn.com/abstract=2765268>.
- Langer, M., König, C. J., & Fiteli, A. (2018). Information as a Double-Edged Sword: The Role of Computer Experience and Information on Applicant Reactions towards Novel Technologies for Personnel Selection. *Computers in Human Behavior*, 81, 19–30. <https://doi.org/10.1016/j.chb.2017.11.036>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesting, A., & Baum, K. (2021a). 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, <https://doi.org/10.1016/j.artint.2021.103473>
- Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021b). Spare Me the Details: How the Type of Information about Automated Interviews Influences Applicant Reactions. *International Journal of Selection and Assessment*, 29(2), 154–169. <https://doi.org/10.1111/ijsa.12325>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J. Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., & Thrun, S. (2011). Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)* (pp. 163–168). IEEE. <https://doi.org/10.1109/IVS.2011.5940562>
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Mantel, S. (2014). No Reason for Identity: On the Relation between Motivating and Normative Reasons. *Philosophical Explorations*, 17(1), 49–62. <https://doi.org/10.1080/13869795.2013.815261>
- Mantel, S. (2015). Worldly Reasons: An Ontological Inquiry into Motivating Considerations and Normative Reasons. *Pacific Philosophical Quarterly*, 98(S1), 5–28. <https://doi.org/10.1111/papq.12094>
- Mantel, S. (2018). Determined by Reasons. *Routledge*. <https://doi.org/10.4324/9781351186353>
- Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>
- McKenna, M. (2012). Conversation and Responsibility. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199740031.001.0001>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*, 22(2), 103–115. <https://doi.org/10.1007/s10676-019-09519-w>
- Mele, A. R. (2021). Direct Versus Indirect: Control, Moral Responsibility, and Free Action. *Philosophy and Phenomenological Research*, 102(3), 559–573. <https://doi.org/10.1111/phpr.12680>
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)* (pp. 36–42).
- Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), <https://doi.org/10.1177/2053951716679679>
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65, 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008>
- Nissenbaum, H. (1996). Accountability in a Computerized Society. *Science and Engineering Ethics*, 2(1), 25–42. <https://doi.org/10.1007/bf02639315>

- Noorman, M. (2020). Computing and Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>
- Pereboom, D. (2014). Free Will, Agency, and Meaning in Life. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780199685516.001.0001>
- Perel, M., & Elkin-Koren, N. (2016). Accountability in Algorithmic Copyright Enforcement. *Stanford Technology Law Review*, 19, 473–533. <https://doi.org/10.2139/ssrn.2607910>
- Polanyi, M. (1966). *The Tacit Dimension*. Routledge and Kegan Paul.
- Potochnik, A. (2007). Optimality Modeling and Explanatory Generality. *Philosophy of Science*, 74(5), 680–691. <https://doi.org/10.1086/525613>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (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* (pp. 1135–1144). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
- Rosen, G. (2003). IV-Culpability and Ignorance. *Proceedings of the Aristotelian Society*, 103(1), 61–84. <https://doi.org/10.1111/j.0066-7372.2003.00064.x>
- Rudy-Hiller, F. (2018). The Epistemic Condition for Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/>
- Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5. <https://doi.org/10.3389/frobt.2018.00015>
- Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 1–28. <https://doi.org/10.1007/s13347-021-00450-x>
- Scanlon, T. M. (2008). *Moral Dimensions*. Harvard University Press. <https://doi.org/10.4159/9780674043145>
- Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to Expect from Opening up ‘Black Boxes’? Comparing Perceptions of Justice Between Human and Automated Agents. *Computers in Human Behavior*, 122. <https://doi.org/10.1016/j.chb.2021.106837>
- Schmidt, E. (2018). Normative Reasons for Mentalism. In Kyriacou, C., & McKenna, R. (Eds.), *Metaepistemology: Realism and Anti-Realism* (pp. 97–120). Palgrave Macmillan. [https://doi.org/10.1007/978-3-319-93369-6\\_5](https://doi.org/10.1007/978-3-319-93369-6_5)
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). <https://doi.org/10.1109/iccv.2017.74>
- Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121(3), 602–632. <https://doi.org/10.1086/659003>
- Shoemaker, D. (2012). Blame and Punishment. In Coates, D. J., & Tognazzini, N. A. (Eds.), *Blame: Its Nature and Norms* (pp. 100–118). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199860821.003.0006>
- Shoemaker, D. (2013). On Criminal and Moral Responsibility. *Oxford Studies in Normative Ethics*, 3, 154–178. <https://doi.org/10.1093/acprof:oso/9780199685905.003.0008>
- Shoemaker, D. (2015). Responsibility from the Margins. *Oxford University Press*. <https://doi.org/10.1093/acprof:oso/9780198715672.001.0001>
- Smith, A. M. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics*, 115(2), 236–271. <https://doi.org/10.1086/426957>
- Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, 24(1), 62–77. <https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Sterz, S., Baum, K., Lauber-Rönsberg, A., & Hermanns, H. (2021). Towards Perspicuity Requirements. In *IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 159–163). IEEE. <https://doi.org/10.1109/REW53955.2021.00029>
- Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48, 1–25.
- Strevens, M. (2017). How Idealizations Provide Understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Essays in Epistemology and the Philosophy of Science* (pp. 37–49). Routledge.
- Talbert, M. (2019). Moral Responsibility. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/>
- Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, 74(4), 905–916. <https://doi.org/10.2307/1954312>

- van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). Moral Responsibility and the Problem of Many Hands. *Routledge*. <https://doi.org/10.4324/9781315734217>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ix005>
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 842–861. <https://doi.org/10.2139/ssrn.3063289>
- Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.
- Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, 24(2), 227–248. <https://doi.org/10.5840/philtopics199624222>
- Zarsky T. (2013). Transparency in Data Mining: From Theory to Practice. In Custers, B., Calders, T., Schermer, B., & Zarsky, T. (Eds.), *Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases* (pp. 301–324). Springer. [https://doi.org/10.1007/978-3-642-30487-3\\_17](https://doi.org/10.1007/978-3-642-30487-3_17)
- Zimmerman, M. J. (1997). Moral Responsibility and Ignorance. *Ethics*, 107(3), 410–426. <https://doi.org/10.1086/233742>
- Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, 32(4), 661–683. <https://doi.org/10.1007/s13347-018-0330-6>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.