Philosophy & Technology (2022) 35:88
https://doi.org/10.1007/s13347-022-00577-5

RESEARCH ARTICLE

AI, Opacity, and Personal Autonomy

Bram Vaassen[1,2]

Received: 15 September 2021 / Accepted: 22 August 2022/ Published online: 22 September 2022

Abstract

Advancements in machine learning have fuelled the popularity of using AI decision algorithms in procedures such as bail hearings, medical diagnoses and recruitment. Academic articles, policy texts, and popularizing books alike warn that such algorithms tend to be *opaque*: they do not provide explanations for their outcomes. Building on a causal account of transparency and opacity as well as recent work on the value of causal explanation, I formulate a moral concern for opaque algorithms that is yet to receive a systematic treatment in the literature: when such algorithms are used in life-changing decisions, they can obstruct us from effectively shaping our lives according to our goals and preferences, thus undermining our autonomy. I argue that this concern deserves closer attention as it furnishes the call for transparency in algorithmic decision-making with both new tools and new challenges.

1 Introduction

Artificial intelligence provides an increasingly popular way of streamlining important and sometimes life-changing decision procedures. Bail hearings (e.g., Feller et al., 2016), medical diagnoses (e.g., Rajkomar et al., 2018; Esteva et al., 2019), and recruitment procedures (e.g., Heilweil, 2019; Van Esch et al., 2019) often rely on such automatized components. Their growing popularity is at least in part due to technical improvements in the design of machine learning algorithms, which drastically increased the amount of data that can be processed, as well as the rate at which it is processed. Crucially, such deep learning algorithms do not only use preprogrammed rules to provide outcomes. They also hone their procedures through

✉ Bram Vaassen
bram.vaassen@umu.se

[1] Department for Historical, Philosophical, Religious Studies Umeå University, Umeå, Sweden

[2] Department of Philosophy, Rutgers University, New Brunswick, USA

numerous trial-and-error runs. Many of these AI decision algorithms are *opaque* even when they are reliable: they might deliver the right results, but they do not provide users or affected parties any insight as to how they came to produce those results.[1]

Many authors treat opacity as one of the central threats posed by our increasing reliance on AI. Calls for transparency in automatized decision-making are commonplace in policy texts (e.g., HLEG, 2019; Information Commissioner's Office, 2021), academic literature (see Floridi et al., 2018), and popularizing books (e.g., O'Neill, 2016; Eubanks, 2018), and are motivated in a variety of ways. I argue that there is a distinct concern with opaque algorithms that has not received a systematic treatment in the literature: opacity can undermine the users' autonomy by hiding salient pathways of affecting the algorithm's outcomes. Importantly, even sufficiently reliable and fair[2] algorithms can pose such a threat to autonomy. Consequently, this threat can remain even if other worries about opaque algorithms are successfully resolved and addressing this threat will require a certain degree of transparency in decision-making.

The text is structured as follows. First, I discuss previous motivations of the call for transparency, as well as some criticism on transparency demands (Section 2). I then provide a causal account of opacity and transparency, and argue that it reveals how opacity can undermine our autonomy. (Section 3). Before concluding, I discuss some practical consequences of opacity's threat to our autonomy (Section 4). In particular, I discuss the impact of viewing transparency issues through the lens of our autonomy on (i) technological obstacles to delivering the right degree of transparency, (ii) prospects for legally entrenching adequate transparency demands, and (iii) weighing the advantages and drawbacks of transparent decision-making.

2 The Call for Transparency

Many participants in the explainable AI (XAI) debates call for transparency, explainability, explicability, or simply for less opacity. The general idea is that users of AI algorithms are entitled to explanations of their outcomes.[3] Quite often, AI algorithms are opaque in the sense that such explanations are not available to

[1] I'll use "users" in a broad sense here, to cover both those who employ and those who develop the algorithms.

[2] In what follows, I will use "fair" to mean "not biased in a problematic way" and assume that this notion of fairness is sufficiently well-understood. This terminological shortcut serves to avoid difficult questions such as what counts as (problematic) bias and whether or not opaque decision algorithms can be "fair" in broader sense.

[3] Some authors provide more fine-grained notions in this area as well, such as "traceability" and "simulatability" (e.g., Lipton, 2018). Although I am convinced that the causal account I will present in Section 3 can capture most of these notions as well, I cannot argue for it here and the arguments presented below go through regardless. Our goal in this text is not to capture all the nuances in XAI terminology, but rather to pick out a moral concern with decisions that are not backed by explanations.

all stakeholders. This opacity can have different sources. Sometimes institutions or corporations fail to communicate when they rely on AI systems or on how these systems work. Alternatively, those who are entitled to transparency might lack the expert knowledge to understand the explanations at hand. Finally, due to the increasing reliance on machine learning, even the most expertly trained humans might fail to grasp the algorithm in full detail.[4] Our primary concern here is not what makes AI algorithms opaque, but rather why we should want them to be transparent.

Transparency is treated as an important requirement for ethical AI implementations. In a survey of the literature on sustainable AI, Floridi et al. (2018) report that "in all [texts surveyed], reference is made to the need to *understand* and *hold to account* the decision-making processes of AI." (p. 700, emphasis in original).[5] The interest in transparency is by no means merely academic. It shows up in both policy texts and popularizing texts as well. For example, the EU guidelines for trustworthy AI specify that at a crucial component of the "transparency" they demand is to "require[] that the decisions made by an AI system can be understood and traced by human beings." (HLEG, 2019, p. 18) and the recent European Commission's AI act proposes that "a certain degree of transparency should be required for high-risk AI systems" (Council of the European Union, 2021, p. 30). Finally, O'Neill (2016, p. 31) lists opacity, damage and scale as the three essential features of algorithms that qualify as "weapons of math destruction." There appears to be broad agreement that transparency and opacity carry substantial moral weight.

This idea certainly has intuitive appeal. Receiving decisions without explanations can be frustrating and scary (cf. O'Neill, 2016, p. 29). Empirical research confirms that our trust in decisions, actions and outcomes increase when we are provided with a plausible explanation (e.g., Herlocker et al., 2000; Symeonidis et al., 2009; Holzinger et al., 2020),[6] and earlier research in psychology suggests that there is a distinct pleasure associated with grasping explanations (Gopnik, 1998). Transparency comes with certain practical advantages as well. Our ability to assess the reliability and fairness of algorithms improves when we grasp the explanations for their outcomes (Kim et al., 2016; Gkatzia et al., 2016; Biran and McKeown, 2017; Doshi-Velez et al., 2017) and improving an algorithm is probably easier when one knows how it works (cf. HLEG, 2019, p. 18).[7] By contrast, opacity often compounds the negative impact of inadequate outcomes. For example, it is harder to question the result of a process if one cannot point the finger at where it went wrong. Consequently, it is harder to negotiate the outcomes of opaque algorithms; their decisions are more likely to go unchallenged when mistaken and those responsible are less

[4] See Burrell (2016), Weller (2019), and Walmsley (2020) for more on these distinctions in opacity.

[5] I am assuming here that understanding a decision-making process requires a certain amount of transparency, but it is important to note that there certainly isn't a one-to-one correspondence between them. More on this later.

[6] Although there are definitely other factors in play as well. See for example Krügel et al. (2022) and Jauernig et al. (2022).

[7] Though see Weller (2019, Section 3.1) for examples suggesting that transparency can lead us to overascribe reliability.

likely to be held to account (cf. O'Neill, 2016; Wachter et al., 2017; Floridi et al., 2018; Walmsley, 2020).

Even so, the call for transparency is not without its critics. Zerilli et al. (2019) maintain that non-AI decision algorithms, such as humans or bureaucratic systems, can be equally complex and mysterious in their actual implementation. We rely on judges and committees to make decisions without having any deeper understanding of how their brain functions.[8]London (2019) emphasizes that common medical interventions often involve mechanisms that are not fully understood by patients, practitioners or researchers. By demanding full transparency of AI decision algorithms, we seem to be holding them to a higher standard than we apply to non-AI decision makers or mechanisms. Reports explaining decisions on job applications never specify the neural details of the committee members, but provide us with very coarse-grained information: "Candidate A's references are unconvincing," "Candidate B lacks the relevant work experience," etc. Such explanations are typically located at the *intentional* level: they concern the beliefs committee members have and how they integrate with their goal of getting the right candidate. An adequate standard of transparency for AI systems would thus similarly demand only restricted access to the workings of the system.[9]

Exactly how the demand for transparency should be restricted is a tricky question. It probably will not do to simply demand transparency at the intentional level, as many maintain that current artificially implemented algorithms do not have such a level (cf. Penrose, 1989; Dreyfus, 1992; Wakefield, 2003). In this article, I identify a distinct moral concern with opaque decision algorithms. Even in cases where an opaque algorithm fulfills its task excellently, and its results are treated as negotiable, its opacity can threaten the autonomy of the parties affected by its outcomes. Getting clear on opacity's threat to autonomy will not only provide extra motivation for demanding transparency, it will also help to determine an adequate standard for transparency. Given that such a threat to our autonomy can occur regardless of whether the decision algorithm is executed by an AI system or by a human, motivating the demand for transparency with its relevance to autonomy would not require a double standard of the kind Zerilli et al. (2019) worry about.

3 Transparency and Autonomy

In order to get clear on the tight relation between transparency and the autonomy of the affected parties, it will help to be more precise on what opacity and transparency actually consist of. I take the causal explanation literature to provide a natural account of transparency and opacity, as well as an intuitive explanation of their relation to autonomy. Let us start with the account of transparency and opacity.

[8] Lipton (2018) and Cappelen and Dever (2021) make similar points in passing.

[9] Though see Günther and Kasirzadeh (2021), who argue that a double standard would in fact be justified.

3.1 The Causal Account

Champions of transparency maintain that users are entitled to an explanation of algorithm outcomes. When an algorithm decides that our job application does not make the cut, we are entitled to knowing why it didn't. A standard way of explaining why events occur is by referring to their causes (cf. Lewis, 1986; Woodward, 2003). The drop in oil prices can be explained by the decreasing demand, because the decrease in demand caused the drop in prices. Similarly, my not getting the job can be explained by my lack of convincing references, if their poor quality caused my application to be rejected. According to the resulting picture, we can characterize opacity and transparency as follows:[10]

> Opacity An AI system is opaque to A to the extent that A is not in a position to grasp the causal explanations of its outcomes. Transparency An AI system is transparent to A to the extent that A is in a position to grasp the causal explanations of its outcomes.

These characterizations need some unpacking. First and foremost, we need to know what a causal explanations are, and what it means to grasp them. Most off-the-shelf accounts from the philosophical literature will do for our purpose, but we will take a broadly interventionist approach as a starting point here. According to such accounts, *X* causally explains *Y* if and only if bringing about changes in *X* while leaving other causes unmeddled with correlates with changes in *Y*, or with the chance of *Y* occurring (e.g., Woodward, 2003, p. 59).[11] For example, if changing my references while leaving all else in my application fixed correlates with my success in getting the job, then my references causally explain my not getting the job. Simplifying quite a bit, we can say that to grasp a causal explanation of *X* is to know what caused *X* and to understand how changes in those causes correlate with changes in *X*.[12]

Second, it is worth emphasizing that such accounts of causal explanation are extremely non-committal about the physical or technical implementation of such counterfactual patterns. For example, as long the right pattern of counterfactual dependence holds between the quality of reference letters and the probability of getting the job, it does not matter how the importance of good reference letters is encoded in the system. In fact, it might very well be that there is no explicit mention of reference letters and what makes them convincing in the code of the algorithm.

[10] Recently, Páez (2019) argued that causal explanation is ill-suited to capture understanding of AI algorithm outcomes. I cannot address his concerns here, but see Erasmus et al. (2020) and Erasmus and Brunet (2022) for a response. See also Frigg and Reiss (2009) for arguments in favor of non-exceptionalism about computer-run algorithms and simulations.

[11] Wachter et al. (2017) also focus on the importance of counterfactuals for transparency, but eschew overtly causal interpretations of these counterfactuals.

[12] See Strevens (2013) for a more detailed account of grasping causal explanations. For our purposes, it is important that understanding how *X* and *Y* correlate does not require full understanding of the mechanism that makes them correlate. I can understand that changes in the position of the volume switch are proportional to changes in the volume output, without understanding the fine details of the relevant mechanism connecting the switch to the volume output.

For many current AI algorithms, it is exceedingly likely that such information is encoded implicitly rather than explicitly (cf. Burrell, 2016, pp. 8–10). According to the broadly counterfactual accounts of causation, being merely implicitly encoded is not an obstacle for being causally effective. As we shall see in Section 4, this feature will make such accounts quite suitable for providing explanations of AI behavior at the right level.

A third point worth elaborating on is that opacity and transparency are matters of degree. One might understand how some outcomes of an AI system are produced, but not how others are. Similarly, one might know some of the causal factors contributing to outcome *X* without knowing all of them. Crucially, not all causal explanations will help us *understand* the outcome, nor will all causal explanations be of interest to us. For example, an overly inclusive causal explanation might confuse us, and thus provide us with no more grasp of how the outcome was produced. On the other end of the spectrum, the explanation "your submitting an application caused you to get a rejection letter" will be of no interest because it provides too little information. As we shall see in Section 4, getting clear on the relation between transparency and autonomy can help us get clear on the degree of transparency required in a given situation, as well as on how to compare the relevance of different causal explanations.

One final point before continuing. Readers familiar with the XAI literature might classify our definitions as focusing on so-called post hoc explanations. This label would be somewhat misleading. Causal explanation can be available *before* the fact in hypothetical form, and events that never occurred can be causally explained in a similar way. For example, we can say "if candidate A were to apply this year, she would not get the job because she has not finished her studies," thus providing a hypothetical causal explanation for an event that may or may not occur in the future. We are often interested in exactly such hypothetical causal explanations. A future applicant is well-advised to ask "if I were to apply, which factors will play a causal role in the outcome of my application?" As will become apparent in the examples to follow, the right degree of transparency will often require that we grasp such hypothetical causal explanations. Our account of transparency focuses on explainability, but this does not restrict us to explainability after the fact.

3.2 Causal Explanation and Autonomy

Starting from our account, the demand for transparency translates into a demand for causal explanations. We can now ask what the value of causal explanation is: why do we or should we want causal explanations of AI algorithm outcomes? Recent work on causal explanation points towards an intuitive answer. Philosophical and empirical research has converged on the thesis that we are particularly interested in causal explanations because they provide us with reliable means to affect and predict our surroundings (cf. Lombrozo, 2011; Hitchcock, 2012). For example, knowing that my subpar references caused me to not get the job allows me to predict that similar jobs will be unavailable to me unless I fix my references. It also provides me with an effective strategy to improve my chances of getting a similar job: improving my references. In effect, opaque decision algorithms hide effective strategies of affecting and predicting their outcomes from the affected parties. Conversely, transparent decision algorithms can enable us to undertake action and affect future outcomes.

In the XAI literature, this action-enabling potential of transparent decision algorithms often goes unmentioned. For example, a recent meta-study of over 100 texts on XAI by Langer et al. (2021), lists twenty-eight desiderata found in the literature but makes no mention of how opacity can undermine the ability of affected parties to effectively influence the outcomes according to their goals.[13] Similar remarks apply to a recent review of thirty-four AI ethics documents published by civil society, the private sector, governments, intergovernmental organizations, and multi-stakeholder organizations by Fjeld et al. (2020).[14]Wachter et al. (2017) is an exception to this pattern and contains a section on transparency as enabling users to alter outcomes. However, Wachter et al. do not explain why our ability to affect such outcomes is important.

This matter deserves closer attention. Our ability to reliably affect and predict the outcomes of decision algorithms can carry substantial moral weight. When it comes to pivotal decisions, such as whether I get a certain job, whether I am allowed into certain college, or whether I get an enormous insurance premium, my not being privy to what factors into these decisions undermines my ability to effectively shape my life. As rational planning agents, we attempt to influence such weighty decisions by preparing accordingly. We study to get the right degree, we select extra-curricular activities that are taken to sharpen the relevant skills, we drive carefully, and so on.[15] Our ability to shape our lives according to our plans and desires and to pursue our goals through deliberate actions, i.e., our ability to be *autonomous* agents, is severely undermined when we are denied information about what factors into life-changing decisions. There are at least three reasons for taking opacity's threat to our autonomy seriously. First, resolving the other worries concerning transparency does not automatically resolve the autonomy worry. Second, the autonomy worry comes with significant backing from moral philosophy. Third, undermining autonomy undermines responsibility. Let us discuss these points in turn.

First of all, the autonomy worry hones in on a feature that is tightly related to the opacity of decision algorithms. In principle, an opaque decision algorithm could be sufficiently reliable, fair, and trusted. Trust can be due to the recommendation of a trusted authority, and reliability and debiasing are eventually just a question of tweaking the algorithm. Certainly, tweaking the algorithm will be harder when we don't have the slightest idea how it works, and in any real-life scenario the programmers would require a minimum of transparency to get going, but there is nothing that *in principle* stands in the way of an algorithm fulfilling its function perfectly

[13] They do list several texts that focus on the autonomy of those who deploy the algorithms, but none that raise a similar worry for the affected parties (Langer et al., 2021, p. 7). As we shall see in Section 4, this focus on users rather than affected parties reverberates in the European Commission's AI act.

[14] The results of these meta-studies are in tension with Selbst and Barocas (2018) claim that transparency's action-enabling potential "dominates" the XAI literature (p. 1120). They refer to Wachter et al. (2017), who indeed dedicate a section to this topic, but further evidence of such dominance is scant.

[15] Though one would hope that we drive carefully for other reasons as well.

without the relevant stakeholders having knowledge of what produces its outcomes. The outcomes could even be treated as negotiable by allowing users to double-check the outcomes using another decision algorithm, be it a human or an artificial one.[16] Such double-checking can also be used as a basis for holding those who deployed or developed the algorithm accountable. By contrast, our lack of knowledge of how to *affect* the outcomes is integral to an AI system being opaque to us. According to the causal account proposed above, opacity and a lack of knowledge about how to manipulate are definitionally inseparable.

The upshot is that even if an opaque algorithm manages to tick all the other boxes, it can still threaten our autonomy. Suppose for example that all applications for government jobs go through the GOV-1 decision algorithm. For the purpose of our example, it matters little how GOV-1 was trained, but we can suppose that GOV-1 is the most reliable system to select government job candidates. All GOV-1 requires to decide who is the right candidate for the job is an accurately filled out questionnaire for each applicant. Careful analysis has demonstrated beyond reasonable doubt that no team of humans or competing artificial algorithms (or any combination thereof) will select a more viable candidate for any government position than GOV-1. We can assume further that applicants have a right of redress by demanding their applications be considered by another (non-AI) system, the government is held accountable for any mistakes by GOV-1 and we can assume trust in the government is strong enough to engender trust in GOV-1 as well. Unfortunately, it is unclear how GOV-1 weighs the information provided in the questionnaire. As a prospective applicant, I have no idea which competences I should acquire in order to become desirable, or even eligible, for a government job.[17]

The opacity of GOV-1 thus hides salient ways of shaping our lives. Perhaps my goal of becoming a government employee requires me to focus more on getting into an international exchange program than on getting high grades. Perhaps I should not have given up my role in the student union to take an extra math credit. These are relatively simple examples, but AI systems can pick up on patterns that are very hard for us to detect. Consequently, GOV-1 might value features that one would assume to be completely irrelevant for being a good candidate. In so far as GOV-1 is opaque to me, I have no way of knowing, and no way of properly weighing my professional goals against any other goals I might have in life: building a family, maintaining friendships, learning to play the sitar… More generally, withholding explanations of decisions amounts to withholding information about how to affect these decisions. When opaque algorithms are used in life-changing decisions, they thus obstruct us from effectively shaping our lives according to our preferences. In short, opaque algorithms can undermine our autonomy, even when they respect other requirements such as reliability, fairness, accountability, and negotiability.

[16] Wachter et al. (2017, p. 98) make a similar suggestion to address cases where the demand for transparency conflicts with trade secrets.

[17] See O'Neill (2016, Ch. 6), Van Esch et al. (2019) and Heilweil (2019) for discussions of actual AI recruitment systems such as Hirevue, ZipRecruiter and Kronos.

A second reason for focusing on opacity's threat to autonomy is that the moral import of autonomy is well-discussed in the philosophical literature. Autonomy takes center stage in several ethical theories. In areas varying from foundational deontology (e.g., Kant, 1993) and utilitarianism (e.g., Mill, 1999), to more applied work on education (e.g., Haji and Cuypers, 2008; Thorburn, 2014), bioethics (e.g., MacKay & Robinson, 2016), political theory (e.g., Raz, 1986), and free will (e.g., Ismael, 2016) authors agree that autonomy carries moral weight. The importance of autonomy is picked up in other discussions within AI ethics (e.g., Kim et al., 2021), and legal theory as well (e.g., Marshall, 2008; McLean, 2009). The call for transparency can draw support from all of these fields.

This appreciation of autonomy's moral importance gave rise to many accounts of what it precisely consists in. It is a further question whether our use of "autonomy" here overlaps or coincides with how it is standardly used in the literature. A detailed comparison will unfortunately have to wait for some other occasion, but here are two remarks on the subject. First, autonomy is taken to be some brand of self-determination and there is growing attention for the fact that self-determination requires a long-term control over one's life and plans (e.g., MacIntyre, 1983; Raz, 1986; Atkins, 2000; Bratman, 2000; 2018; Ismael, 2016). Second, even authors whose accounts would not appear to mesh well with our usage here acknowledge that cases like GOV-1 need to be taken into account. For example, Christman (1991) defends a broadly internalist view of autonomy, according to which our desires should fit our rational beliefs and the rationality of a belief does not require its being true. So in principle, prospective applicants can rationally believe that studying political science rather than physics will help their chances, even if it does not. Even so, Christman accepts that rationality requires a somewhat reliable connection to reality:

> One is autonomous if one comes to have one's desires and beliefs in a manner which one accepts. If one desires a state of affairs by virtue of a belief which is not only false but is the result of distorted information given to one by some conniving manipulator, one is not autonomous just in case one views such conditions of belief formation as unacceptable (subject to the other conditions I discuss) (Christman, 1991, p. 16).

Generally speaking, it would be surprising if no lack of information could undermine our autonomy. Based on these two observations, one can reasonably expect that autonomy as the ability to shape one's life will correspond sufficiently with common usage in ethics.

A third reason to focus on the threat to autonomy is that autonomy strongly correlates with responsibility. Generally speaking, undermining an agent's autonomy relative to an outcome undermines responsibility relative to that outcome as well. This is because responsibility for an outcome requires a reliable causal connection between the agent's intention to reach or avoid the outcome and the outcome in fact being reached or avoided (cf. Björnsson and Persson, 2012; 2013; Grinfeld et al., 2020; Usher, 2020). If candidate A intends to get a government job, but has no idea how they should polish their competences in order to qualify, the reliability of the correlation between her intending so and her achieving her goal should be expected to decrease. Generally speaking, not knowing how to achieve a goal makes it less likely that you achieve it. When opaque algorithms undermine our autonomy, they also undermine our responsibility for the outcome.[18]

In conclusion, it appears that opacity can undermine autonomy and autonomy has moral value. Even if we set aside the above-mentioned connection with trust, fairness, reliability, accountability, negotiability, and a primitive preference for explanation, demands for transparency can still be grounded in a requirement to respect personal autonomy.

4 Transparency in Practice

Grounding the demand for transparency in autonomy requirements sheds new light on some familiar issues in XAI. I elaborate on three of these here: (i) the question how much transparency is desirable and whether this degree is technically attainable, (ii) how to entrench the demand for transparency legally, and (iii) whether transparency conflicts with other desiderata decision procedures. Let us discuss these in turn.

4.1 Delivering the Right Degree of Transparency

The connection between transparency and autonomy provides guidance in deciding how transparent decision algorithms ought to be. We want to know how differences in the input correlate with differences in the output. It is well-received in both the philosophical and the computer science literature that knowledge of such higher-level regularities does not presuppose knowledge of the finer details of the system (e.g., Dennett, 1971; 1991; Newell, 1982; Campbell, 2008). Establishing which level of explanation is appropriate for the relevant input-output correlation will no doubt be an arduous task that requires different strategies for different cases, but there is a general point to be made here as well. In order to increase or maintain our autonomy, we want to know which changes in the input robustly correlate with certain changes in the output. Some patterns of correlation will be too fragile to be of genuine interest. Perhaps having a twitter handle without numerals increases your chances with 0.02% if you are a Caucasian woman with a law degree from a foreign university, but will have no effects in any other circumstances. Other patterns will be crucial knowledge for future applicants. Perhaps the only way to get a government job without a university degree is when you score above 150 on the IQ test and can demonstrate a staunch unwillingness to believe conspiracy theories. That is to say, in most circumstances, a university degree is a condition *sine qua non*. Such robust correlations are worth knowing.

It is unlikely that knowledge of such robust correlations will require full physical details or design details, but it is also unlikely that the required explanation will always be found at the intentional level. This is not only because AI systems might

[18] See Baum et al. (2022) for a worked out argument along these lines with a focus on users in the loop rather than affected parties.

typically lack such a level altogether (cf. supra), but also because the most robust patterns might not be found at this level. To see this, consider the human case of implicit bias. While such biases are not typically implemented at the intentional level, they can still make for robust correlations between certain input features and outputs. For example, committees with racist implicit biases might systematically review candidates with foreign-sounding names unfavorably, without those biases being manifested at the intentional level (cf. Bertrand and Mullainathan, 2004). The upshot is that there is no fixed level that provides the right level of transparency. Instead, we should focus on finding those correlations that are robust across the scenarios in which the algorithm is to be applied.

The good news is that such input-output difference-making transparency is easier to achieve than full physical, design or algorithmic transparency. Full physical transparency would require a grasp of the workings of the system in all its physical details down to the electrons making up the hardware. Achieving such transparency is of course very difficult, but also not very useful. Full algorithmic transparency requires a grasp on the mathematical details of the algorithms encoded in the AI system and design transparency requires an engineering perspective on how such a system can be developed. Attaining either of these is taken to be extraordinarily difficult as well, and demanding it might even be in tension with copyright laws (cf. Burrell, 2016; Ananny & Crawford, 2018; Wachter et al., 2017). By contrast, strategies for attaining transparency in input-output correlations are available. For example, Tubella et al. (2019) propose to test AI system's conformity with moral norms by using a "glass-boxing" method to check whether the input-output correlations of AI algorithms conforms to moral norms. A glass box is built around a system by checking its inputs and outputs. In Tubella et al.'s proposal, this glass box should detect input-output pairs that violate moral norms, but attaining input-output transparency requires less work. All the glass box should do is report the correlations between differences in inputs and differences in output. As the glass box is "built around" the AI system, this method would not require us to "open up" the algorithm that is being tested. Standard causal extraction algorithms, such as those developed by Pearl (2000) and Spirtes et al. (2000), can be used to acquire causal information on the basis of the correlational data gathered via glass-boxing. As Woodward (2003) bases his account of causal explanation on the structural equation models provided by Pearl (2000) and Spirtes et al. (2000), the causal account of transparency and opacity we proposed in Section 3 naturally fits this technical approach.[19] Several other strategies that do not require "opening up" the black box have been developed (cf. Guidotti et al., 2018; Ustun et al., 2019; Belle and Papantonis, 2021). Insofar as these strategies can in fact provide us with robust patterns of correlations between inputs and outputs, they can help safeguard user autonomy.[20]

[19] Admittedly, such extraction algorithms require a certain causal structure to the data set in order to be successful. This need not be a problem as we typically have access to such minimal information. We know, for example, that for each individual case, the inputs cause the outputs, and not vice versa.

[20] This is not to say that other concerns with AI opacity, such as assessing reliability and fairness will not require us to open the black box. As I have argued above, these are separate issues

Undivided optimism about such "forensic" approaches would be premature. First of all, these approaches all omit details about the actual process leading up to the outcome when providing potential explanations. Such neglect of detail is necessary to provide explanations that are understandable for human agents, and, so the worry goes, there is a real risk that the omitted details are in fact crucial to the true causal story of how the decision was in fact produced (cf., Rudin, 2019). In the worst case, this lack of detail may make for mistaken explanations altogether, taking mere correlation for causation. While this is a real risk, it is worth noting that this risk is by no means unique to complex algorithms. It is well-recognized that the explanations of any event will require us to omit enormous amounts of details that, strictly speaking, contributed to its coming about (e.g., Loewer, 2007; Ney, 2009). There is always a risk of omitting crucial details and focusing on irrelevancies. If we are to dismiss forensic approaches in general just because they are at risk of getting things wrong, we would be holding AI decision algorithms to a far higher transparency standard than we hold any other kind of explanation. The relevant question is whether these algorithms can reliably provide the right explanations in the contexts where they are implemented. Whether this can be done is an outstanding empirical question. If we are to use AI to make decisions that significantly impact our lives, this question needs to be investigated in-depth on a case-by-case basis.

Rudin (2019) formulates a more distinctive challenge for current forensic approaches as well. Their ability to provide counterfactual information about what would have changed the outcomes of a particular procedure often relies on the notion of a "minimal" difference to the input (e.g., Ustun et al., 2019). The underlying idea is that users and affected parties are mostly interested in how they could have changed the outcome with as little effort as possible: I'd rather drastically improve my chances at getting a job by taking an evening course in business English than by taking a full law degree, for example. However, what counts as a "minimal" difference for the affected party is highly dependent on their context. Perhaps one's family context makes it easier to get a higher-paying job than to move to another ZIP code in order to get a mortgage. It thus probably won't do to provide a one-size-fits-all explanation for each outcome that just mentions the one difference-maker the algorithm deems "minimal." A wider variety of realistic strategies to affect the outcome should be made available, such that the affected party can select the strategy that fits them best, if any. Whether providing such a smörgåsbord of options is practically feasible is, again, an open question.

There are ongoing attempts to make provide explanations of black box outcomes without "opening" black boxes. In order for these "forensic" strategies to safeguard autonomy, they need to reliably provide accurate explanations that fit the needs and practical perspective of the affected parties.

4.2 Legally Entrenching Transparency Demands

Transparency demands play a central role in recent attempts to regulate the use of AI. However, the current formulations of these demands will not suffice to respect the autonomy of affected parties, as they do not provide the them with a right to explanation. Focusing on the European Union General Data Protection Regulation (GDPR) (Council of the European Union, 2021) and the European Commission's AI act (Council of the European Union, 2016), I briefly elaborate on some of the obstacles for legally entrenching the relevant transparency demands. There is some hope that its link with autonomy can provide the call for transparency with extra legal backing, but there is still a long way to go.

There are two salient challenges with legally entrenching a right to explanation that will sufficiently respect the autonomy of the affected parties. First of all, if the AI algorithms in use are under protection of proprietary laws, the developers and deployers can maintain that any demand to disclose the workings of the algorithms conflicts with their right to intellectual property. True, the "forensic" tools discussed above could allow for the relevant causal knowledge without detailed knowledge of the algorithmic implementations. But making the case that the relevant causal knowledge can be acquired without the detailed knowledge that is defended by proprietary laws is likely to be an uphill battle. If it has to be shown for each demand for an explanation that the required explanation would not be in conflict with trade secrets, the affected parties are likely to find themselves *de facto* disenfranchised with regard to their right to explanation. Wachter et al. (2017, p. 98) make note of this challenge and suggest that external auditing mechanisms can be employed in such cases, but while such mechanisms could help to enforce the right of redress and to justify accusations of bias, they would not provide explanations, and thus not provide affected parties with action-enabling information.

Second, our previous discussion on finding the right level of explanation further complicates the picture. Any outcome will have many explanations that are of no practical use to the affected parties. For example, one causal explanation of why I did not get a government job might be that my application activated node 64879508, which activated node 0540324875, but inhibited node 45009783245. And, as it happens, such an activation pattern provides a negative outcome. Without any background knowledge about the role of these nodes within the system, this explanation is of no help to me, even if it is factual. The upshot is that the right to *an* explanation will not suffice. The explanation needs to fit the explanatory needs and perspective of the affected party. While there are several excellent guides on how to provide explanations that are fitting in this sense (e.g., Information Commissioner's Office, 2021), it remains an open question whether demands for a "fitting" translation can be legally entrenched. Even if a right to *an* explanation can be made to go through, this further dimension needs to be kept in mind.

These difficulties with entrenching an adequate right to explanation transpire in recent attempts to regulate AI use. Wachter et al. (2017) point out that the GDPR fails to provide affected parties with a right to explanation of individual decisions, but instead delivers a rather vague right to be informed of (i) whether an AI is used, and (ii) what "logic" underlies the algorithm. Moreover, even this watered down right appears to apply only when the decision is based *solely* on automated algorithms (Art.22(1)). Which means that even a minimal involvement of a human agent in the process can relieve the users of any obligation to provide information about the algorithm (cf., Wachter et al., 2017, p. 78). I refer the reader to the original text for a detailed discussion of their evidence. Suffice it to say that the kinds of explanations that can sustain our ability to effectively pursue our life plans would not be protected by the GDPR alone.

More recently, the European Commission's AI act proposes to impose transparency requirements on precisely those algorithms with outcomes that have significant impacts on the affected parties, such as decisions on legal status, access to education or contractual relations (Annex III).[21] However, as argued by Fink (2021), the proposal, if accepted, would only demand transparency towards the *users* of AI, and not towards the affected parties.[22] The only obligation towards the affected parties would be to inform them of the fact that an AI decision algorithm has been used to reach a conclusion. The AI act thus does little to legally entrench our autonomy in the face of decision algorithms that are entirely opaque to us.

There is some hope that a right to an adequate explanation of automatized decisions can be derived from more general rights. Fink (2021) points out that article 41.(2c) of the European Charter of Fundamental Rights concerns exactly the obligation of the administration to "give reasons for its decisions." However, this article focuses on a right to good *administration*, where "administration" is taken to refer "the institutions, bodies, offices and agencies of the Union." Therefore, it is of no help to those who seek reasons for decisions made by insurance companies, banks, or private educational facilities. Moreover, the articles in the charter are explicitly restricted to cases where the administration are "implementing Union Law" (Art.51), which puts a further burden on affected parties to argue that the governmental decision they demand explained counts as an implementation of Union Law. There are several legal cases where a plaintiff's reliance on the right to good governance as spelled out in Art.41 is deemed illegitimate because the governmental decisions in question were not considered to count as implementations of Union Law.[23] So, even when it concerns governmental decisions, Art.41(2c) of the charter provides only a limited right to explanation.

In light of the precarious position of the right to explanation in current legislation, the link between transparency and autonomy can perhaps be of help. There are at least some promising signs to be found here, as the notion of personal autonomy plays a central role in a variety of legal frameworks. For example, the European Court for Human Rights has relied on the notion of personal autonomy in several rulings,[24] and some legal scholars maintain that the very right to personal autonomy is enshrined in the European Convention of Human Rights (ECHR) (e.g., Marshall,

[21] Note though, that the demands on AI decisions in "non-high risk" contexts are limited to mere guidelines that are adopted on a voluntary basis. This means that affected parties must be able to convincingly argue that the decisions affecting them count as "high risk." This risks putting an undue burden of proof on victims who do not have the means to litigate in grey area cases.

[22] This is despite the fact that the document explicitly distinguishes between users and affected persons (p. 30). This oversight is reminiscent of the general oversight in the texts reviewed by Langer et al. (2021), which focused on the autonomy of the users, but made no mention of the autonomy of the affected parties (see Section 3).

[23] See for example (Raad van State, 2017, Section 17). Similar arguments are found in the Belgian Council of State decisions on cases 232.758 (29.10.2015), 233.512 (19.01.2016), and 238.292 (23.05.2017).

[24] See Koffeman (2010) for an overview

2008) and the US Constitution (e.g., McLean, 2009). If respecting personal autonomy requires providing adequate explanations for potentially life-changing decisions, then the right to personal autonomy requires a right to explanation.

However, there are significant obstacles on this route towards a right to explanation as well. Neither the US Constitution or the ECHR explicitly mention autonomy. Instead, arguments for the right to autonomy based on these texts tend to go via related notions, such as dignity (ECHR, Art.1) and privacy (ECHR, Art.2). Consequently, the jurisprudence at the ECHR is equivocal on how much weight is to be attached to autonomy. Some judges rely on autonomy as a guiding notion to interpret these foundational legal texts, whereas others take the right to autonomy to follow from texts themselves.[25] It certainly would not hurt to have more concrete formulations of the importance of autonomy and a legal right to explanation that is adapted to the current rise of automated decision-making available in our legal frameworks.

While the call for transparency has inspired the notion to play a central role in legal documents focusing on AI, these documents do not guarantee transparency towards affected parties whose lives are affected by the outcomes of automatized decision procedures. The link between transparency and autonomy provides a promising extra tool for legally entrenching a right to explanation. Even so, there is still plenty of legislative work to be done before we can feel safe that such a right is legally protected.

4.3 Downsides of Transparency

Even if transparency has a distinct moral value in serving autonomy, it should not be pursued at all costs. There may be any number of reasons to trade off transparency for other goods. I focus here on three potential drawbacks of providing the degree of transparency that is required to bolster autonomy.

First, providing information of the robust correlations that allow us to affect the outcomes of decision algorithms might increase the advantage of those who have easier access to the difference-making features. For example, if attending expensive private universities increases one's chances of getting a government job, it might be overall justifiable to not divulge this fact. As it becomes easier to control the outcomes in fair ways, it will become easier to control the system in unfair ways as well. Implementing transparency requirements will require consideration of this fact (cf. Weller, 2019, Section 3.2).

Second, it is a well-received truism that otherwise reliable indicators can become unreliable once it is publicly known that they are used as indicators (Campbell, 1979; Goodhart, 1984). For example, if word gets out that GOV-1 treats participation in debate club as a big plus, this might cause students to attend debate club merely to improve their chances at a government job, rather than to develop the relevant skills that debate club is supposed to foster, like absorbing and structuring information, and presenting clear arguments. The influx of new students with less

[25] See Koffeman (2010, pp. 8–9) for discussion.

intrinsic motivations to acquire these skills might drastically decrease the reliability with which debate clubs cultivate these skills in their participants, hence making participation in debate club a less reliable indicator. Making criteria used in life-changing decisions accessible to affected parties might similarly change the reliability of those same criteria.

Third, building on work by O'Neill (2002), Nguyen (forthcoming) recently argued that transparency can improperly limit the kinds of reasons that feature in decision-making. Forcing experts to make their reasons for certain judgments accessible and understandable for non-experts, risks forcing them to limit their reasons to the kind of reasons that non-experts are sensitive to. This would in effect make it impossible to invoke reasons that require advanced expertise to appreciate. While Nguyen does not focus on automatized decision-making, his arguments appear to transfer quite easily to the automatized case. In fact, one of the oft-cited reasons for using automated decision algorithms is that they appear to discover patterns that are not readily appreciable by human observers. If we only use transparent algorithms, we might come to see that some unexpected or simply intractable reasons feature in the production of the decision. In the long run, this might incite us to use less reliable mechanisms that only employ reasons and inferences that are tractable to us.

So even though transparency can help bolster autonomy, this does not mean that it is to be pursued at all costs. Relatedly, it is worth emphasizing that the argument presented here does *not* establish that transparency or autonomy are *intrinsically* valuable. The argument I have provided in favor of transparency is clearly restricted to its instrumental value. I have argued that transparency carries moral weight as a means for supporting user autonomy. The argument leaves it open whether or not (i) transparency is intrinsically valuable, (ii) autonomy is intrinsically valuable, or (iii) transparency has instrumental value beyond its contribution to autonomy. Colaner (2021) recently argued in favor of (i), and many of the works referenced throughout this text provide evidence that (ii) and (iii) are true as well. Even so, our central argument goes through even if (i)–(iii) turns out to be false: opaque decision algorithms can undermine our autonomy by hiding salient pathways of affecting their outcomes. If the broad consensus that autonomy carries moral weight is correct, this means transparency is worth demanding.

5 Conclusion

There are several reasons to demand that impactful decision algorithms are transparent. This is true regardless of whether they are implemented by humans or AI systems. Previous research indicated that transparency is conducive to negotiability and accountability, helps fine-tune reliability and avoid bias, and that we are more likely to trust the results of algorithms if we understand what causes their outcomes. Building on recent work about the value of explanation, I have argued that a lack of transparency can also undermine the autonomy of the affected parties. In both the human and the AI case, we are interested in knowing the robust patterns of correlation that allow us to reliably affect and predict their outcomes. Such knowledge can play an integral part in planning and shaping our lives as rational, self-determining agents.

This perspective on transparency furnishes XAI debates with both new tools and new challenges. On the one hand, calls for transparency can draw on established work in moral philosophy and legal texts that emphasizes the importance of personal autonomy. Moreover, focusing on the autonomy of affected parties can guide us in deciding what kinds of explanations we should demand. On the other hand, it appears that resolving previous concerns about opacity will not automatically address the threat to the autonomy of the affected parties, and the kinds of explanations required to respect our autonomy can be hard to come by. It might not require us to open the black box, but it does require us to take into account the perspectives of rational planning agents with life-goals and dreams.

Acknowledgements This text greatly benefited from comments by audiences at Umeå University, the #FrAIday lecture series, the WGPT at KU Leuven, the Umeå-Göteborg practical philosophy workshop at Göteborg University, and the 'Governing AI panel' at MANCEPT. Special thanks to Wouter Vaassen en Koen Willebrords for helping me navigate the legal texts.

Author Contribution Single-authored paper

Funding Open access funding provided by Umea University. Deutsche Forschungsgemeinschaft (Grant No. FOR 2495) Swedish Research Council 2019-06381

Data Availability Not applicable

Code Availability Not applicable

Declarations

Conflict of Interest The author declares no competing interests.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Atkins, K. (2000). Autonomy and the subjective character of experience. *Journal of Applied Philosophy*, *17*(1), 71–79.

Baum, K., Mantel, S., Speith, T., & Schmidt, E. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy and Technology*, *35*(1), 1–30.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 39.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013.

Biran, O., & McKeown, K.R. (2017). Human-centric justification of machine learning predictions. In *IJCAI*, (Vol. 2017 pp. 1461–1467).

Björnsson, G., & Persson, K. (2012). The explanatory component of moral responsibility. *Noûs*, *46*(2), 326–354.

Björnsson, G., & Persson, K. (2013). A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research*, *87*(3), 611–639.

Bratman, M. (2018). *Planning, time, and self-governance: Essays in practical rationality*. Oup USA.

Bratman, M.E. (2000). Reflection, planning, and temporally extended agency. *Philosophical Review*, *109*(1), 35–61.

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

Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90.

Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, *18*(1), 426–445.

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. New York: Oxford University Press.

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy*, *21*(1), 1–24.

Colaner, N. (2021). Is explainable artificial intelligence intrinsically valuable? *AI and Society*, 1–8.

Council of the European Union. (2016). General Data Protection Regulation. https://gdpr-info.eu/. Accessed 21 April 2022

Council of the European Union. (2021). European Council AI Act. https://artificialintelligenceact.eu/the-act/. Accessed 23 April 2022

Dennett, D.C. (1971). Intentional systems. *Journal of Philosophy*, *68*, 87–106.

Dennett, D.C. (1991). Real patterns. *Journal of Philosophy*, *88* (1), 27–51.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S.J., O'Brien, D., Scott, K., Shieber, S., Waldo, J., Weinberger, D., & et al. (2017). Accountability of AI under the law: *The role of explanation. Berkman Center Research Publication, Forthcoming*.

Dreyfus, H.L. (1992). *What computers still can't Do: A critique of artificial reason*. Cambridge: MIT Press.

Erasmus, A., & Brunet, T.D.P. (2022). Interpretability and unification. *Philosophy and Technology*, *35*(2), 1–6.

Erasmus, A., Brunet, T.D.P., & Fisher, E. (2020). What is interpretability? *Philosophy and Technology*.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Feller, A., Pierson, E., Corbett-Davies, S., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post* 17.

Fink, M. (2021). The EU artificial intelligence act and access to justice. *EU Law Live*.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication* (2020-1).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., & et al (2018). AI4People — An ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.

Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese*, *169*(3), 593–613.

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. arXiv:1606.03254

Goodhart, C.A. (1984). Problems of monetary management: the uk experience. In *Monetary theory and practice* (pp. 91–121). Springer.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*(1), 101–118.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J.F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42.

Günther, M., & Kasirzadeh, A. (2021). Algorithmic and human decision making: for a double standard of transparency. *AI & Society*, 1–7.

Haji, I., & Cuypers, S.E. (2008). Authenticity-sensitive preferentism and educating for well-being and autonomy. *Journal of Philosophy of Education*, *42* (1), 85–106.

Heilweil, R. (2019). Artificial intelligence will help determine if you get your next job. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen. Accessed 12 Feb 2021

Herlocker, J.L., Konstan, J.A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241–250).

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

HLEG, A. (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines\#Top. Accessed 28 Jan 2021.

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: the system causability scale (scs). *KI-Künstliche Intelligenz*, 1–6.

Information Commissioner’s Office. (2021). Explaining decisions made with AI. https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/. Accessed 28 March 2021.

Ismael, J. (2016). *How physics makes us free*. USA: Oxford University Press.

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. *Philosophy and Technology*, *35*(1), 1–25.

Kant, I. (1993). *Grounding for the metaphysics of morals: With on a supposed right to lie because of philanthropic concerns*. Hackett Publishing Company.

Kim, B., Koyejo, O., Khanna, R., & et al. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS* (pp. 2280–2288).

Kim, T.W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, *70*, 871–890.

Koffeman, N. (2010). (the right to) personal autonomy in the case law of the european court of human rights (nota opgesteld ten behoeve van de staatscommissie grondwet). *(The right to) personal autonomy in the case law of the European Court of Human Rights (nota opgesteld ten behoeve van de Staatscommissie Grondwet)*.

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? humans trust untrustworthy ai-advisors for ethical decisions. *Philosophy and Technology*, *35*(1), 1–37.

Langer, M., Oster, D., Speith, T., Kästner, L., Baum, K., Hermanns, H., Schmidt, E., & Sesing, A. (2021). What do we want from explainable artificial intelligence (xai)? ? a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, *296*, 103473.

Lewis, D.K. (1986). Causal explanation. In D. Lewis (Ed.) *Philosophical papers*, (Vol. 2 pp. 214–240). Oxford University Press.

Lipton, Z.C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Loewer, B.M. (2007). Mental causation, or something near enough. In B.P. McLaughlin J.D. Cohen (Eds.) *Contemporary debates in philosophy of mind* (pp. 243–64). Blackwell.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551.

London, A.J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21.

MacIntyre, A.C. (1983). *After virtue: A study in moral theory*. University of Notre Dame Press.

MacKay, D., & Robinson, A. (2016). The ethics of organ donor registration policies: Nudges and respect for autonomy. *American Journal of Bioethics*, *16* (11), 3–12.

Marshall, J. (2008). *Personal freedom through human rights law?: Autonomy, identity and integrity under the European convention on human rights*. Brill.

McLean, S.A. (2009). *Autonomy, consent and the law*. Evanston: Routledge.

Mill, J.S. (1999). *On Liberty*. Broadview Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18* (1), 81–132.

Ney, A. (2009). Physical causation and difference-making. *British Journal for the Philosophy of Science*, *60*(4), 737–764.

Nguyen, C.T. (forthcoming). Transparency is surveillance. *Philosophy and Phenomenological Research*.

O'Neill, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown Publishing Group.

O'Neill, O. (2002). *A question of trust: The BBC Reith lectures 2002*. Cambridge: Cambridge University Press.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Penrose, R. (1989). *The Emperor's New Mind*. New York: Oxford University Press.

Raad van State. (2017). nr. 237.630 in de zaak a. 213.945/ix-8508. http://www.raadvanstate.be/Arresten/237000/600/237630.pdf#xml=http://www.raadvanstate.be/apps/dtsearch/getpdf.asp?DocId=36730&Index=c%3a%5csoftware%5cdtsearch%5cindex%5carrets%5fnl%5c&HitCount=1&hits=219d+&04252620222717. Accessed 24 Apr 2022.

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., & et al (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1* (1), 1–10.

Raz, J. (1986). *The morality of freedom*. New York: Oxford University Press.

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

Selbst, A.D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham. L Reviews*, *87*, 1085.

Spirtes, P., Glymour, C., & Richard, S.N. (2000). *Causation, prediction and search*. Cambridge: Mit Press.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515.

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2009). Moviexplain: a recommender system with explanations. In *Proceedings of the third ACM conference on Recommender systems* (pp. 317–320).

Thorburn, M. (2014). Values, autonomy and well-being: Implications for learning and teaching in physical education. *Educational Studies*, *40*(4), 396–406.

Tubella, A.A., Theodorou, A., Dignum, V., & Dignum, F. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. arXiv:1905.04994

Usher, M. (2020). Agency, teleological control and robust causation. *Philosophy and Phenomenological Research*, *100*(2), 302–324.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10–19).

Van Esch, P., Black, J.S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222.

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.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GPDR. *Harv. JL & Tech.*, *31*, 841.

Wakefield, J.C. (2003). The Chinese room argument reconsidered: Essentialism, indeterminacy, and strong AI. *Minds and Machines*, *13*(2), 285–319.

Walmsley, J. (2020). Artificial intelligence and the value of transparency. *AI and Society*, 1–11.

Weller, A. (2019). Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning* (pp. 23–40). Springer.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology*, *32*(4), 661–683.