Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–March:1–12

DOI: 10.1177/20539517231222872
journals.sagepub.com/home/bds

Sage

Algorithmic decision-making: The right to explanation and the significance of stakes

Lauritz Aastrup Munch[1], Jens Christian Bjerring[1] and Jakob Thrane Mainz[1]

Abstract

The stakes associated with an algorithmic decision are often said to play a role in determining whether the decision engenders a right to an explanation. More specifically, "high stakes" decisions are often said to engender such a right to explanation whereas "low stakes" or "non-high" stakes decisions do not. While the overall gist of these ideas is clear enough, the details are lacking. In this paper, we aim to provide these details through a detailed investigation of what we will call the "Simple Stakes Thesis." The Simple Stakes Thesis, as it will turn out, is too simple. For even if the stakes associated with a specific one-off decision are low—and hence does not engender a right to an explanation—such decisions may nevertheless form part of a high stakes *pattern* or *aggregate* of decisions. In such cases, we argue, even a low stakes decision may engender a right to explanation. Not only does this show that the right to explanation is more demanding than so far recognized but it also shows that the stakes thesis is significantly harder to apply in practice.

Keywords

Algorithmic decision-making, black box algorithms, right to explanation, stakes

Introduction

In response to an increasing reliance on artificial intelligence in making decisions that affect people, many have advocated for a right to explanation. According to a widespread view, people have a right to explanation when they are subjected to one-off high stakes algorithmic decisions. As Coyle and Weller put the idea,

> "there is a growing demand to be able to "explain" machine learning (ML) systems' decisions and actions to human users, particularly when used in contexts where decisions have *substantial implications for those affected* and where there is a requirement for political accountability or legal compliance" (Coyle and Weller, 2020: 1433; our *italics*).

Standard examples of high stakes algorithmic decisions concern decisions about credit lending, hiring, university admission, healthcare prioritization and diagnostics, social security payments, and sentencing (Cao and Yousefzadeh, 2023; Coyle and Weller, 2020; Kempt et al., 2022; Wong et al., 2023). While virtually everyone accepts that there is a right to explanation in such high stakes cases, some go even further and say that a decision's status as "low stakes" is sufficient for there *not* to be a right to explanation.

Here are some representative expressions of these sentiments:

> "[…] *If the stakes are low, the right to explanation does not apply* […]. *If the stakes are high* […], the right to explanation rules out opaque decision procedures […]" (Vredenburgh, 2022: 18; our *italics*).

> "The use of AI for low-risk decisions carries little, if any, moral hazard and so does not require explanations" (von Eschenbach, 2021: 1620).

> "There seem to be many implementations of AI in situations of low to no risk (in terms of harm). It is unreasonable that the decisions resulting from AI in these situations should be required to provide explanations" (Robbins, 2019: 507).

[1]Department of Philosophy and History of Ideas, Aarhus Universitet, Aarhus, Denmark

Corresponding author:
Lauritz Aastrup Munch, Department of Philosophy and History of Ideas, Aarhus Universitet, Aarhus, Denmark.
Email: laumu@cas.au.dk

> "[…] [h]igh-stakes decision making and troubleshooting (…) are the main two reasons one might require an interpretable or explainable model" (Rudin, 2019: 1).

Extrapolating from these quotes, at least three distinctive ideas are in play. First, we are presented with the overarching idea that the stakes associated with a decision matter as to whether a right to explanation is engendered for that decision. Second, we are told something about the direction in which stakes matter: namely that *high stakes* engender a right to an explanation; notions such as "significantly affected," "real-life stakes," and "high risk" all mean roughly, we take it, the same as "high stakes." Third, we are told that stakes only matter for explainability insofar as they are sufficiently high; or alternatively, that low stakes decisions do not engender a right to explanation.[1] We will use this bundle of ideas to formulate below what we will call the *Simple Stakes Thesis* about the right to explanation. Clarifying and assessing that thesis is the main aim of the paper.

We proceed in three steps. First, in section II, we offer a reasonably precise account of the Simple Stakes Thesis. Second, in section III, we consider its justification. Inspired by Vredenburgh, 2022 (but see also Wachter et al., 2017 and Taylor, 2023) we attempt to justify the thesis by appeal to self-advocacy and autonomy; but, as we shall see, there are other accounts that might vindicate the significance of stakes as well. Third, in section IV, we critically scrutinize the Simple Stakes Thesis and argue that it remains importantly incomplete. Specifically, while the thesis entails that low stakes algorithmic decisions do not generate a right to explanation, we argue that they can generate such a right in cases where they form *part of a pattern* of decisions whose outcome, when taken as an aggregate, may be of significant importance to individuals. Not only does this demonstrate that the right to explanation is more demanding than previously acknowledged but it also highlights the considerable challenges involved in applying the stakes thesis in practice.

Our conclusions are important for a number of reasons. First, although the literature acknowledges that stakes have a central role to play in carving out the scope of the right to explanation in contexts of algorithmic decision-making, so far only little effort has been made to explain and motivate this idea in detail. Second, our claim that the stakes associated with patterns of decisions—as opposed to only one-off decisions—matter for the scope of the right to explanation serves to align the AI literature on the right to explanation closer with findings in moral and decision theory. While it is widely recognized in these fields that the appropriate context for evaluating a decision problem often extends beyond isolated, one-off decisions (e.g., Dietz, 2023; Parfit, 1984; Stefánsson, 2023; Thoma, 2019), the broader AI debate on the right to explanation has not yet fully explored these insights. This oversight is unfortunate because it blurs the normative importance of explaining and justifying algorithmic decisions that are not, on the standard understanding, high stakes decisions. Finally, the concept of stakes plays a central role in shaping the scope of legal rights in regulatory contexts. For instance, several of the legal rights listed in the GDPR—including rights to explanation—apply only if automated decision-making produces legal effects that "[…] significantly affects [people] […] such as automatic refusal of an online credit application or e-recruiting practices" (recital 71). Along with these lines, by clarifying how and when algorithmic decisions may "significantly affect" people, we can help refine our understanding of the relevant set of legal rights and their potential scope.

Clarifying the simple stakes thesis

Suppose you have your final exam tomorrow. Only a high grade will get you into medical school. You have dreamt of medical school your whole life, so you feel very stressed about the exam. When asked about why you feel stressed, you answer "Because there is so much at stake." What do you mean when you say this? Plausibly, you mean that whether you get a good enough grade means the world to you. Not only will failing to score a high grade have a negative impact on your future career opportunities but it will also come with huge personal costs.

Let us generalize this familiar way of thinking. Consider a simple decision-process with two possible outcomes. We can think of the stakes of the decision as a function of the difference in "choiceworthiness" between these two outcomes. If the "choiceworthiness" of an outcome can be measured numerically, we can express the stakes of a decision as the value you would put on the most preferred outcome minus the value you would put on the least preferred outcome. The higher this value is, the higher are the stakes associated with the decision-process. Of course, some decision processes have more than two possible outcomes. For instance, a number of algorithmic risk assessment tools output risk numerical estimates ranging from 0 to 1 (Patty and Penn, 2023). In such cases, we can express the stakes associated with a decision as a function of the distance between the most choiceworthy and the least choiceworthy outcome of the decision.

This rather thin account strikes us as giving a natural understanding of stakes, and we will adopt it to reason about the stakes of both algorithmic and nonalgorithmic decisions in what follows. The account assumes an independent theory of what makes an outcome choiceworthy. Providing such a theory, however, is philosophically controversial because it is an open question which features make an outcome choiceworthy. *Subjectivists* about well-being may say that what determines the value of an outcome depends on the preferences of the person affected by the outcome (Crisp, 2001). If we pair this subjectivist

view with the idea that stakes matter, the preferences of those affected by algorithmic decisions will hence play a key role in establishing the scope of the right to explanation. To illustrate, compare two persons applying for medical school. While one person has a strong desire to go to medical school and become a doctor, the other is indifferent and only applies because the application process entertains them. Here, the subjectivist may say that the stakes are high only for the first person. And this, in turn, will determine whether a right to an explanation of the admission decision is generated. By contrast, *objectivists* about well-being may say that what determines the value of an outcome depends on objective features of the outcome that are independent of people's actual preferences. While some objectivists may emphasize the role of individuals' preferences in determining the value of an outcome, they need not. In the example above, for instance, an objectivist may locate properties of the outcomes that make the stakes high for both applicants. Such properties may involve certain objective value-enhancing aspects associated with being admitted to medical school, including access to education, the acquisition of skills that benefit others, the opportunity for meaningful employment, and so forth. While various hybrid views of well-being are conceivable as well (e.g., Wenar, 2023), our strategy here will be to rely on intuitive judgments about what determines the values of algorithmic outcomes. As far as we can tell, our conclusions are compatible with different views on what makes an outcome choiceworthy.

Using the conception of stakes as the difference in choiceworthiness between possible outcomes, we can now define:

> **Simple Stakes Thesis**. A subject who is affected by a decision D is owed an explanation of D when and because the stakes associated with D are sufficiently high.

The Simple Stakes Thesis is a threshold view: it says that the stakes of a decision—algorithmic or not—must be "sufficiently" high to generate a right to explanation. It is reasonable to adopt a threshold view in this context because "explanation," in contrast to stakes, is nongradable.

Before getting into the details, let us first see the thesis in action:

> **Loan:** A subject applies for a loan in a bank. The bank deploys a complicated algorithmic system for calculating creditworthiness, and they use the resulting creditworthiness score as a basis for deciding whether the subject is granted a loan.

We take this to be a paradigmatic example of a high stakes algorithmic decision where the subject affected by the decision is entitled to an explanation of the decision. Since it is easy to imagine situations where the value difference in loan outcomes is substantial—cases, for instance, where getting a loan determines a family's opportunity to buy a house—there are clear cases where the stakes associated with individual loan decisions are high. In such cases, as desired, the Simple Stakes Thesis says that the subject is owed an explanation of the algorithmic outcome.

Getting to the details: let us first ask what it means to say that somebody is owed an explanation. Broadly speaking, we can think of explanations as answers to why-questions (Lipton, 2001; Ross, 2023; Vredenburgh, 2022; Woodward, 2019). "Why does she not love me anymore?" and "Why is autonomy valuable?" are both requests for explanations. We are concerned with explanations that target decisions produced by algorithmic systems. Minimally, this involves addressing questions like why a statistical model, in a specific context, produced a particular probabilistic outcome such as an estimation of creditworthiness or a risk score of developing breast cancer. Questions that dig into the technical workings of statistical models do not exhaust what we have in mind here. For the estimates of algorithms are often not identical to the decisions that actually impact people. A decision to grant a bank loan will often rely on a nonalgorithmically determined threshold, which is used to translate an algorithmic scalar estimate of creditworthiness into a binary decision (Beigang, 2022). There might, for instance, be steps in the relevant decision process where humans interpret the algorithm's probabilistic outcomes with a view toward making a final verdict on the bank loan. Since an understanding of such steps in the decision process is often needed for explaining the final decision satisfactorily, we will rely on a broad understanding of what constitutes an algorithmically aided decision. Indeed, while our interest lies with the right to explanation in cases of algorithmically aided decision-making, we do not believe that such a right is unique to the algorithmic context. On the contrary, people plausibly often have moral rights to receive explanations of decisions that are based entirely on human judgments (e.g., Vredenburgh, 2022). This naturally raises the question of whether we should hold algorithmic and human-based decision-making to the same explanatory standards. This is the topic of a lively debate—see, for instance, Günther and Kasirzadeh, 2022, and Zerilli et al., 2019—that we need not, however, weigh in on directly here. So although we focus on algorithmic decision-making, our central results should apply in contexts where human decision-making is part of the process too, and they should be consistent with different ways of understanding the relative explanatory standards we impose on algorithmic versus nonalgorithmic decision-makers.

A second question concerns the types of explanations that individuals are owed due to the right to explanation. Matters are complicated here because different types of explanations can be normatively significant in different contexts, and because these different types of explanations themselves can be subjects of theoretical discussions. Since our main aim is to *explore* the Simple Stakes Thesis, however, we can sidestep some of these thorny issues.

Yet, it is not too controversial to grant that causal explanations will be of particular importance here.[2] Causal explanations aim to provide causal understanding of a particular explanatory target—for instance an algorithmic decision—by citing its causes (Lipton, 2001; Selbst and Powles, 2017; Vredenburgh, 2022). Philosophers disagree over how to best conceptualize causation, but an influential tradition on which we shall rely propose that we analyze causes as a species of *difference-makers* (e.g. Pearl, 2000; Woodward, 2003). To illustrate, suppose we want an explanation of why someone was denied a loan by an algorithmic system, and suppose this explanation cites as a cause the person's low annual income. In giving this explanation, we are implying that *had* the person *not* had a low annual income, they would have been granted the loan. In other words, we imply that income amounts to a difference-maker with regard to the algorithmic decision, and hence that income is at least a partial cause of the relevant decision.

Let us say a bit more about causal explanations in the context of algorithmic decision-making. In what follows, we will focus on causal relations that describe how algorithmic models work. Citing income as a central factor in explaining an algorithmic estimate of creditworthiness is saying something about how the model functions. These sort of internal relations may or may not track the real-world causal relations they sometimes are meant to model. For instance, it may be that there also exists model-external causal relations between income and creditworthiness in the sense that people with higher income have a higher objective chance of paying back their loans—indeed, such objective chances may explain why the algorithm uses income to estimate creditworthiness in the first place. But these model-external relations are not the explanatory targets that we are interested in here.

By holding that the right to explanation sometimes requires causal explanations of specific algorithmic outputs, we may worry whether it is even feasible to provide such explanations. This worry is partly fueled by the infamous *black-box* problem (e.g., Burrell, 2016). In a nutshell, the problem is that many advanced machine learning algorithms—especially deep learning models—operator in manners that are opaque and not easily interpretable, even by their creators. We can observe how these algorithms reliably relate certain inputs with certain outputs, but we cannot easily tell *why* they make the correlations that they do. While we shall not tap into this debate here, we should mention that there is a promising literature on *Explainable AI* (XAI), which is dedicated to developing algorithmic tools and techniques that can make these complex models more transparent and understandable to users, stakeholders, and regulators. More specifically, there is hope that these XAI tools may at least give some of the things that causal explanations typically give us (Divyat et al., 2019; Eoin et al., 2022; Fleisher, 2022; Greta et al., 2022; Keane et al., 2021; Sahil et al., 2020). For instance, *Shapley values* may provide some information about which parts of an input space were of particular importance for generating a specific algorithmic outcome (Bhargava and Gupta, 2022), and *counterfactual techniques* may reveal some of the (nonspurious) counterfactual dependence relations in the algorithmic model that are necessary for identifying difference-makers (for discussion, see Baron, 2023; Beckers, 2022; Buijsman, 2022; Chou et al., 2022). Of course, we are not suggesting that the XAI suite of techniques amounts to a silver bullet when it comes to explaining algorithmic systems, but merely that they show promise in providing at least a partial causal understanding of these systems.

Finally, some question whether it is possible to explain in causal terms what AI systems are doing. While we need not take a strong stance on this issue here, it is worth noting that many at least think that difference-making is central for capturing explanatory relations (see Taylor, 2023 for discussion and further references). We believe that difference-making is central to causality as well, but strictly we only need that difference-making is central to explanations. In the case of counterfactual XAI techniques, for instance, the fact that they give us means for identifying difference-making properties of an algorithm's input space may well be enough to explain the algorithm's behavior—bracketing the further question of whether these properties also track a causal structure.

While the remarks above hopefully serve to narrow down the kinds of explanations that we will focus on, let us remind ourselves, going forward, that our main focus is not so much on giving a detailed account of the kinds of explanations that we are entitled to in high stakes algorithmic contexts. Rather, we are mainly concerned with understanding why and when, from a moral standpoint, we should insist on receiving such explanations.

Justifying the simple stakes thesis

We have clarified the elements of the Simple Stakes Thesis and can now ask about its justification. That is, what could justify us in holding that we have a right to a (causal) explanation of algorithmic decisions in high stakes contexts?

A promising answer to this question emphasizes our need to be able to make *informed decisions* in contexts where much is at stake. Vredenburgh calls the capacity to pursue our interests informedly the capacity for *informed self-advocacy* (Vredenburgh, 2022). Informed self-advocacy, she says, "is a cluster of abilities to represent one's interests and values to decision-makers and to further those interests and values within an institution" (Vredenburgh, 2022: 212). Informed self-advocacy is closely intertwined with agency and autonomy, and it emphasizes the significance of having explanations of

algorithmic decisions that directly affect us. For without such explanations, we are unable to adjust our behavior in response to the algorithmic outcome in ways that further our interests and values. For instance, if we do not know which input variables of an algorithm were causally responsible for giving us a low credit score, we will not know how to manipulate these variables—typically through modifying our behavior to align with patterns that the algorithm rewards—to better serve our interests. Likewise, without explanations, our abilities to contest and scrutinize algorithmic decisions are severely hampered (Wachter et al., 2017), which makes it hard for us to engage in rational means-ends reasoning. Being able to explain algorithmic decisions thus has great instrumental value by enabling us to optimize our chances of successfully maneuvering various complex (social) environments.

Since the values associated with self-advocacy and autonomy typically increase as the stakes associated with the algorithmic decision increase, we hence have a neat justification for why stakes matter for the right to explanation. Consider again Loan but assume this time that the applicant's central life plans are seriously advanced if the loan is granted, but severely hindered if denied. Clearly, if the applicant does not understand why the algorithm gave him a particular credit score, he cannot easily further his interests and values. That is, he will not be able to modify his behavior to boost his chances of getting a high algorithmic credit score and thus advance his life plans. Nor will he be able to scrutinize the algorithm's decision procedures to possibly contest its evaluation. So without relevant explanations, the applicant is essentially unable to act in ways that will further his central life plans, depriving him of the possibility of exercising self-advocacy and autonomy.

Given that the values associated with explainability in this way correlate with the stakes of the algorithmic decision, we also have a neat justification for why low stakes decisions do *not* engender a right to explanation—at least not through the broadly normative mechanisms that we are after. If the stakes associated with an algorithmic decision are low, we know that there is not much difference in choiceworthiness between the various outcomes of the decision. And since part of what makes an outcome choiceworthy concerns the extent to which it promotes our interests and values, the outcomes associated with low stakes decisions hence do not differ much in their abilities to promote these interests and values. Accordingly, if low stakes decisions only have a negligible impact on our abilities to exercise self-advocacy and autonomy, it is only natural that these decisions on their own do not engender a right to explanation. So we have a ready justification for why we are not owed an explanation for algorithmic decisions, which pertain to particular CAPTCHA-tests, particular song recommendations on Spotify, particular video suggestions on Youtube, and so on.

But could a claim not be made that there is a right to explanation even in low stakes decision contexts, albeit only to an explanation of a different, less capacious kind? While it is unclear that the verb "explain" admits of degrees, perhaps explanations can be degreed as a function of their underlying properties. It is well-known that explanations can vary in their good-making properties by displaying different levels of specificity, accuracy, localization, and completeness. For instance, an explanation that merely indicates which features matter for an algorithmic outcome is less specific than an explanation that also gives precise estimates for how these features matter. If we acknowledge such variations in explanations, we may then argue that even low stakes contexts engender a right to explanation. It is just that the explanations in these contexts are less capacious than those in high-stakes contexts when judged by their specificity, accuracy, localization, or completeness.

There is certainly no need to rule out this option, and perhaps it is better to conceive of the right to explanation as a phenomenon that can be graded in accordance with underlying good-making properties such as those mentioned above. For our purposes, however, it makes sense to stay in the more explicitly binary vocabulary. Suppose with us that informed self-advocacy and autonomy explain why stakes are important for grounding a right to explanation. If low stakes decisions only have a minimal impact on self-advocacy and autonomy, then there is no obvious reason to insist that explanations—capacious or not—are owed in low stakes decision contexts. For in these contexts, an explanation of the decision does not contribute to promoting or preserving self-advocacy and autonomy.[3]

At this point, one might also observe that self-advocacy and autonomy are not the only—or even the most compelling—justifications for the idea that stakes matter for the right to explanation. We agree with this point and believe that we should be *pluralists* about what moral factors ground the right to explanation. For concreteness, we restrict our attention to the autonomy-based way of grounding the right to explanation, but it is instructive to compare our approach with some alternatives before moving on.

Lazar (forthcoming) says that a right to explanation can be grounded in the need for *legitimate authority*. Central to his view is that algorithmic systems may create or restructure power relations between people, and that such power relations must be constrained in certain ways to be legitimate. One such constraining mechanism may be achieved through providing explanations of algorithmic outcomes: in doing so, we foster the accountability of those exercising power through algorithmic means and ensure that they can be subjected to public scrutiny (ibid.). But what then explains the normative significance of legitimate authority? Lazar gestures at two answers. First, he offers a *self-determination rationale:* being subjected to decisions that are left unexplained undermines self-determination (ibid.,

section III). We take this idea to be a close cousin of the autonomy rationale we have offered above. Second, he offers a *relational equality rationale* (ibid.). The idea is that if some person exercises unchecked power over another, then they do not relate as equals. One way to maintain unchecked power is to obscure it from those it governs —by, for instance, withholding information about its operations and rationale. Since explanation offers a means to clarify the use of power to those affected by it, a right to explanation may then be necessary for fostering equal relations. For our purposes, what matters is that Lazar's story can arguably help justify something like the Simple Stakes Thesis. One obvious reason is that the self-determination rationale is partly grounded in the kind of autonomy-based considerations that we used to support the significance of stakes. Things are more complicated with Lazar's equality rationale. On the face of it, people can fail to relate as equals independently of what is at stake. People may, for instance, fail to relate as equals by *not regarding* others as equals, and this can happen independently of practical stakes. If so, then stakes are not directly relevant for the equality rationale. Nevertheless, Lazar does seem to believe that stakes are important. He writes that the stakes associated with the exercise of power might matter to its legitimacy: "[w]henever the stakes are comparably high, and power is used to govern, the legitimacy and authority standards should have some force" (ibid.). So our exploration of the significance of stakes seems largely complementary to Lazar's approach, even if his account does not fully bottom out in concerns for autonomy.

Purves and Davis' (2022) offer a trust-based account of why algorithmic decisions ought to be explainable. Central is the idea that opacity can undermine trustworthiness because opacity prevents trustors from monitoring whether decision-makers operate in normatively acceptable ways. As a way of grounding the Simple Stakes Thesis, however, a trust-based account is not obviously applicable because it is unclear that trusting an algorithm is sensitive to stakes. For example, even if you have no understanding of an algorithm, which is used solely for making low stakes decisions, questions may still arise about whether it is rational for you to trust the algorithm. As such, it seems that questions about trust in algorithms can arise irrespective of the stakes associated with the decisions that they make. Of course, Purves and Davis might deny that questions about trust in algorithms are relevant in low stakes contexts. In that case, their account would effectively boil down to the Simple Stakes Thesis. But the claim that trust in algorithms is irrelevant in low stakes contexts requires significant argumentation. So it is at least not clear that justification for the idea that stakes matter for the right to explanation depends on issues pertaining to trust.

We might also attempt to justify the significance of stakes through considerations of fairness. In passages, Vredenburgh seems to suggest that the right to explanation is needed for establishing *procedural fairness* (2022: 210). Plausibly, the idea is that opaque decision-making rules out a fair decision-procedure. Assuming that fairness complaints hold little weight against algorithmic systems that only impact people minimally, this perspective may help us motivate something akin to the Simple Stakes Thesis. As above, we need not dispute that fairness considerations can provide a rationale for the Simple Stakes Thesis. For the thesis to matter and being worth exploring, it need not be uniquely justified in concerns for autonomy.

Why the simple stakes thesis is incomplete

Part of what makes the Simple Stake Thesis attractive lies in its simplicity. By considering the stakes associated with individual algorithmic decisions, we can decide, on a case-by-case level, whether the decision ought to be accompanied by an explanation. While this might seem to suggest that the thesis is easy to apply in practice, the story, however, is more complicated.

We have seen that the Simple Stakes Thesis entails that low stakes algorithmic decisions do not engender a right to explanation. But low stakes algorithmic decisions, as we shall explain now, may form *part of a pattern* of decisions whose outcome, when taken as an aggregate, may matter greatly to individuals. So while it may be true that low stakes decisions do not engender a right to explanation, when taken in isolation, we want to argue that they can engender such a right when taken as an aggregate. Since the Simple Stakes Thesis fails to consider such aggregated effects on decision outcomes, the thesis is incomplete.

To help clarify this aggregation idea, consider first the following example from Parfit who owes it to Glover and Scott-Taggart (1975):

> "The Drops of Water. A large number of wounded men lie out in the desert, suffering from intense thirst. We are an equally large number of altruists, each of whom has a pint of water. We could pour these pints into a water-cart. This would be driven into the desert, and our water would be shared equally between all these many wounded men. By adding his pint, each of us would enable each wounded man to drink slightly more water—perhaps only an extra drop. Even to a very thirsty man, each of these extra drops would be a very small benefit. The effect on each man might even be imperceptible." (Parfit, 1984: 78).

Each individual act of contributing a pint of water to the communal water-cart has a minimal impact on any individual thirsty man. In isolation, each pint's contribution to quenching someone's thirst is negligible. For the thirsty man, that is, there is only a small variation in the choice-worthiness linked to each individual act of either contributing or not contributing a pint of water to the cart. As such, the stakes associated with each individual act are low for the

thirsty man. Yet, the individual acts form part of a significant and nontrivial pattern of decisions whose outcome, when taken as an aggregate, matters greatly to the thirsty man. In this sense, when we take into consideration such aggregated effects, the individual acts may matter greatly to the thirsty man.

We encounter this kind of aggregation phenomenon in many contexts. In election contexts, each individual vote decision carries only little, if any impact on who will be elected for presidential office. But each individual decision nevertheless forms part of a significant pattern of decisions whose outcome, when taken as an aggregate, matters tremendously for who will be elected. Likewise, although each individual decision to fly on holiday twice a year may only have a negligible impact on the global CO2-emission spreadsheet, the individual decisions still form part of a pattern of decisions whose outcome, when taken as an aggregate, do have significant impact on this spreadsheet.

While we trust that the general aggregation phenomenon is clear enough, there is so far no consensus about what explains it (Nefsky, 2015). Somehow, it seems, people are moved by *pattern-based reasons* when they make decision:

> "Many of us feel that our reasons for or against acting in particular ways can depend not only on the features of our own actions, but also on the features of the larger patterns of action in which we would or could be participating: […] we feel we have not only *act-based* reasons, but also *pattern-based* reasons. […] Even though our action might not in itself make a morally significant difference, we may still feel that we should participate in good patterns, and should not participate in bad patterns. For example, you might think that if it would be ideal for us to elect a particular political candidate, then that fact gives you some reason to do your part and vote for that candidate." (Dietz, 2023: 131)

The philosophical questions surrounding these issues are intriguing, but for our purposes, we only need the idea that decisions with negligible outcomes can matter because of aggregation effects. And most people seem to accept this idea, irrespective of what exactly explains these effects. Moving forward, we will thus accept not only that the outcome of a specific one-off decision does not inherently give us reason to demand an explanation of that decision—as the differences in choiceworthiness between its outcomes may be minimal. But we will also accept that we can acquire such reasons if we find that the specific one-off decision forms part of a larger pattern with significant outcomes.

Let us now explore a few examples that illustrate these ideas:

Price Discrimination. A consumer visits a webshop. Opaque algorithms are employed by the webshop to create consumer profiles, enabling the webshop to personalize the prices of different products based on the algorithms' predictions of each consumer's willingness to pay. The prices displayed on the webshop vary marginally for different consumers based on their different profiles, but consumers do not understand how the profiling works. The consumer ends up buying a product that would have been offered to him at a slightly lower price, had the algorithm profiled him differently.

Advertising. A social media platform employs an algorithm to display customized advertisements to users. The algorithm utilizes user profiling to optimize the likelihood of user engagement with the ads. However, the user lacks understanding of the algorithm's classification criteria that they are subjected to. Over time, as the algorithm continues to present targeted ads, it succeeds in influencing the user to develop a preference for luxurious specialty coffee.

Demonetization. A content creator uploads videos to a platform with the expectation of earning a small revenue by allowing ads to be embedded in the videos. The monetization of the video is determined by an algorithm, which uses opaque criteria to assess the video's suitability for generating revenue. Each video earns a negligible revenue through its lifecycle on the platform, but since the content creator uploads many videos, the total revenue is significant.

Navigation. A frequent traveler relies on Google Maps as a trusted navigation tool in his daily commute. Unbeknownst to the traveler, a bug in the code of the route planner causes it to consistently select routes that are slightly longer than necessary. As a result, the traveler experiences a slight increase in travel time, with each journey taking a few seconds longer than it would if the bug did not exist.[4]

These cases all involve one-off algorithmic decisions where the difference in choiceworthiness between the different algorithmic outcomes is negligible for most individuals. In Navigation, for instance, we may think of the difference in choiceworthiness between the algorithmic outcomes as pertaining to a slight decrease or increase in whatever value is associated with experiences of length of travel time. In Price Discrimination, we may think of these differences as pertaining to a slight monetary benefit. Insofar as the inferior algorithmic outcome, by stipulation, occurs in both Navigation and Price Discrimination, the individuals in both cases experience minor loses or missed benefits as a result of the algorithmic decisions. However, given that the one-off losses and missed benefits are of very little significance for these individuals—indeed, it is hard to see how the outcome of each individual algorithmic decision could significantly impact an individual's ability to exercise

self-advocacy and autonomy—the differences in choiceworthiness between the different algorithmic outcomes are plausibly negligible for them. As such, the stakes associated with any particular one-off instance of these algorithmic decisions can be considered low. Since what we say about Navigation and Price Discrimination hold for the other cases above too, the stakes associated with the relevant one-off algorithmic decisions are hence low. Accordingly, the Simple Stakes Thesis tells us that no right to an explanation is engendered for these types of decisions.

But insofar as the algorithmic decisions continue to impact an individual over a significant number of iterations, the aggregated consequences of the interactions with the algorithms in the cases above will cease to be negligible for the individual. Consider again Navigation, and suppose that the user of the buggy navigation system is a seasoned delivery person who for a decade has relied on the system to calculate several delivery routes a day. For each route planned by the buggy navigation system, a few seconds is added relative to similar routes planned by a nonbuggy system. Let us say that travel time is the relevant unit of measure, and that travel time is something that a delivery person wants to minimize to free up valuable time. We can then associate a certain value X of choiceworthiness with the first (and actual) pattern of algorithmic decisions by adding together all the travel times of the route suggestions by the buggy system. We can also associate a value Y of choiceworthiness with the alternative (and nonactual) pattern of algorithmic decisions that are made up by the nonbuggy system. By adding together all the travel times of the route suggestions by the nonbuggy system, the value for Y will be significantly smaller than the value for X. Over the time span of a decade, that is, the fact that each route planned by the nonbuggy system is slightly quicker will aggregate to a significant decrease in total travel time compared to the buggy system. For illustration, if 20 routes are planned each day for 250 days a year over a 10 years' time period, and each pair of buggy vs. nonbuggy route differ by two seconds travel time, the aggregate difference in choiceworthiness between the patterns X and Y would amount to $((20 \times 250 \times 10) \times t)$ - $((20 \times 250 \times 10) \times (t - 2)) = 100.000$ s. Of course, saving 100.000 s of travel time may seem insignificant when compared to being denied a loan, but it is certainly still not nothing that is at stake here. Accordingly, we can say that the stakes associated with the first pattern X of algorithmic decisions are higher than those associated with the alternative pattern Y of algorithmic decisions, and hence that the stakes associated with the two patterns correspond to the difference in aggregate travel time.

Insofar as the stakes associated with patterns of algorithmic decisions can vary in this manner, it is also clear why having an explanation of a specific such pattern can be significant for an individual. Had the individual in Navigation, for instance, been offered an explanation of the pattern of algorithmic decisions, he could potentially have become aware of the bug in the system. Suppose the bug occurs if the individual accesses Google Maps via a browser but not if he accesses it via a dedicated app. As motivated earlier, we can expect that an appropriate explanation of an algorithm's decision-making should mention factors that significantly affect the algorithm's outcome. In the case at hand, we can imagine that one of these factors concerns whether the navigational system is accessed through a web browser or a dedicated app. By having an explanation of the pattern of algorithmic decisions that makes this difference salient, the individual would have enough information to infer that his travel time is inflated due to the way he accesses the navigational system. And while having this explanation might not matter much for one-off uses of the navigational system, it matters for repeated long-term interactions with the navigational system as motivated above.

Insofar as we grant that we can compare the choiceworthiness of patterns of algorithmic decisions—as we can compare the choiceworthiness of one-off algorithmic decisions—we can apply this line of explanation to the other cases above as well. Take Advertising. In this example, there is also a significant difference between the one-off interaction with an algorithm showing you a luxury coffee ad based on your profile, and the repeated exposure to that specific ad over an extended period of time interacting with the algorithm. While a single exposure to the coffee ad might not impact your coffee preferences much, the repeated exposure to the ad might eventually result in you developing a new and much more expensive coffee preference. So while the stakes may be low with respect to each individual algorithmic decision to display a specific ad, they can become significantly higher when considering patterns of repeated decisions to expose users to the same ad. Indeed, given the financial impact that the aggregated algorithmic decision may have on the individual's life in Advertising—in conjunction with all the other algorithmically generated ads that the individual is presumably exposed to—it is clear why we value being informed about such aggregated decisions. For instance, had the individual known that the algorithm, for purposes of user profiling, utilizes aspects of his psychology and his socioeconomic situation to potentially implant a new preference in him, he could have used that information to make an informed choice about whether to continue interacting with the algorithm or not.

So we trust that it is clear how individuals can care about having an explanation of aggregated algorithmic decisions while not caring (much) about having an explanation of any particular one-off algorithmic decision. Aggregates of algorithmic decisions like the ones above can have significant impact on peoples' lives and their abilities to exercise autonomy and self-advocacy. In this sense, there is no big difference between one-off algorithmic decisions and

patterns of such decisions: both can be associated with levels of choiceworthiness and hence be said to have stakes associated with them. So if we have a right to an explanation of high-stakes one-off algorithmic decisions —and not many seem to disagree with this—we should also have a right to an explanation of high-stakes *aggregated* algorithmic decisions. Yet, since the Simple Stakes Thesis only applies to one-off algorithmic decisions, the thesis cannot explain how these patterns or aggregates of algorithmic decisions can engender a right to explanation. Thus the Simple Stakes Thesis is incomplete.

Of course, we may think that we can understand the Simple Stakes Thesis as applying to aggregated algorithmic decisions as well. After all, we just have to remember that what constitutes a relevant algorithmic decision must countenance aggregations of one-off algorithmic decisions. If we understand quantification over decisions in the Simple Stakes Thesis to include such aggregations, the thesis can thus stand—although it arguably no longer deserves the adjective "simple." We have no quarrel with this way of framing things. Philosophically, what matters is that we can no longer decide if an algorithmic decision should be accompanied by an explanation simply by looking at the stakes of specific one-off algorithmic decisions. Instead we have to ponder complex questions concerning how aggregates of algorithmic decisions may impact peoples' abilities to exercise autonomy and self-advocacy over time.

Granting that we have a right to an explanation of high-stakes aggregated algorithmic decisions, there are questions about what the accompanying explanation should look like. As touched upon above, if various XAI tools can help us shed light on one-off algorithmic decisions, it seems that they can also help us shed light on aggregates of such decisions. To motivate this thought further, consider Price Discrimination, and suppose that the consumer is provided with a *counterfactual* explanation of the algorithm's decisions. A counterfactual explanation would likely indicate how the values of the input variables used to create a consumer profile would have to change for the price of the relevant product to change. For example, the consumer might learn that had his demographic information suggested a lower socioeconomic status, then the algorithm would have predicted a smaller willingness to pay for the product. The consumer can then use this sort of counterfactual information to gain insights into the algorithm's pricing of products and into how he may affect the outcomes of the algorithm's decisions. Given a high enough number of repeated interactions with the algorithm, such information could translate into significant financial benefits. Something similar happens in Advertising. By receiving an explanation of how the algorithm creates a profile of the user, the user can better act in anticipation of how the social media company is trying to influence his online behavior. To be sure, things will get more complicated when we allow aggregations of *different* algorithmic decisions, and when we focus on aggregative mechanisms that do not merely—as our examples suggest—add low stakes-decisions together.[5] Yet, on the surface of it, it at least appears as if XAI tools can play an important role in answering explanatory questions about aggregates of algorithmic decisions.

One may worry, though, that the appeal to XAI methods spawns a *target mismatch* in our proposal. When low stakes algorithmic decisions aggregate, we are interested in understanding the pattern of decisions produced. Yet, the XAI methods discussed above seem to target individual algorithmic decisions, in which case these methods will not give us what we are after when it comes to understanding aggregates. We reply: if the aggregate pattern is indeed produced by mechanisms that operate at the level of individual algorithmic decisions, then understanding how these individual decisions were made will be crucial for understanding how the pattern emerges. If so, then even if XAI tools only target individual algorithmic decisions, they will be vital for explaining how the aggregated pattern of algorithmic decisions came about.

But this answer may raise a new worry. Even if an individual could in principle understand patterns of algorithmic decisions through explaining individual decisions by use of XAI tools, it may be cognitively infeasible to gain understanding in this way. If so, it is hard to see how explanations of aggregates of decisions could serve people's autonomy in practice. One thing to note is that this is in fact a broad-scoped worry about the value of the right to explanation; see Vredenburgh (2022) for extended discussion. While we cannot hope to deal with this worry comprehensively here, we can at least offer two mediating remarks. First, in saying that XAI tools and explanations of individual decisions may serve people's autonomy, we are not ruling out that further interventions may be necessary to foster understanding; as a start, the relevant explanations will have to be tailored to different individuals. Second, in cases where the pattern of interest is composed of *homogenous* algorithmic decisions—i.e., decisions that abide the same algorithmic decision rules—worries about cognitive overload may be less serious. For here understanding of a single low stakes algorithmic decision may translate into understanding of all the decisions in the pattern. But we fully recognize that there is an important practical challenge involved in promoting understanding of patterns that are produced by many decisions via *heterogeneous* decision procedures.

In closing, let us emphasize that our main argument concerns the question of whether explanations are owed for low stakes algorithmic decisions. While our conclusions show that explanations can be owed for such decisions, we can imagine many different institutional setups that could honor this insight in practice. We have focused on XAI methods here, but for all we have said, the best institutional setup might be one where people—*before* they are

subjected to algorithmic decisions—are presented with rule-like or covering law-like explanations of how the relevant algorithms operate. Insofar as such broad-scoped explanations would enable people to understand the kinds of decisions that the relevant algorithms produce—in combination with knowledge of the relevant input parameters to the algorithms—the obligation to provide explanations of said algorithmic decisions could also be fulfilled.

Conclusion

In this paper, we have attempted to specify in detail the widespread idea that the stakes associated with an algorithmic decision can engender a right to explanation. We have argued that the Simple Stakes Thesis should be augmented to take into account both the stakes associated with one-off algorithmic decisions, and the stakes associated with patterns or aggregates of such decisions. This way of scoping the idea that stakes matter for algorithmic decision-making, we claim, fits better with the reasons for why we are concerned with providing explanations of algorithmic decisions in the first place. No doubt there is more to explore regarding the importance of aggregated algorithmic decisions for the right to explanation. But we are happy to leave these tasks for future work. Our main aim here has been to get clearer on the core aspects of the idea that stakes matter for algorithmic decision-making.

If we are right, it is not easy to implement in practice the idea that stakes matter for algorithmic decision-making. For we can no longer solely consider the choiceworthiness of outcomes from one-off algorithmic decisions to determine whether a right to explanation is engendered with respect to such one-off decisions. Rather, to properly determine whether we have a right to an explanation of one-off algorithmic decisions—especially those that we tend to associate with low stakes—we need information about the likely pattern of interactions that people might have with the relevant algorithms over time. Since such patterns can obviously be hard to identify, a refined stakes thesis will be quite difficult to apply in practice. Put differently, while there has been a lot of focus in the literature on simple cases like Loan, our findings show that there is a vast range of ordinary cases that are much more complicated to handle in a stakes-based framework.

Moreover, if we are right, then what matters morally often concerns the stakes associated with patterns of algorithmic decisions. Since these patterns are hard to identify reliably, there are difficult questions about how we should design the institutions that are supposed to ensure that explanations are delivered to the relevant right-bearers. As we have already seen, using the stakes of a specific decision as a criterion for when the right to explanation applies will undersupply explanations. So instead we might opt for the policy that *all* algorithmic decisions should be explainable. The obvious problem is here, of course, that it is costly to provide explanations—at least assuming that one type of explanation will not fit all types of algorithmic decisions—and that we generally only can impose such costs on people when there is an adequate justification for doing so. So an oversupply of explanations will be problematic as well. So a more balanced approach seems preferable. In essence, a balanced approach would focus on excavating the most reliable evidence that we have for determining whether a specific, one-off algorithmic decision is likely to form part of a pattern of decisions that may impact a person significantly. This approach goes beyond simply assessing the choiceworthiness of specific algorithmic outcomes. It also requires that we consider how people engage with a decision-making system over time, as well as how that system interacts with other systems and societal structures to create aggregate effects.

Accordingly, there is a lot to process for proponents of the idea that stakes matter for whether we are owed an explanation of algorithmic decisions. Indeed, you may worry that an idea, which at least initially appeared simple and compelling, is getting too complicated and difficult to formulate precisely. In that case, you may be willing to grant the Simple Stakes Thesis but deny that decisions can aggregate and form nontrivial patterns in the way we have suggested. We agree that we could have done more to convince such "aggregation skeptics," but hopefully we have done enough to convince those skeptics that they owe us a story about why stakes should *not* aggregate in the way that examples such as Prince Discrimination and Advertising suggest.

Acknowledgments

The authors thank the audience at the Nordic Network for Political Theory (Aarhus 2021), an audience at the Uehiro-CEPDISC workshop (Copenhagen 2021) as well as three anonymous reviewers for helpful comments and suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work in this paper has been generously supported by the Carlsberg Foundation (grant number: CF20-0257).

ORCID iD

Lauritz Aastrup Munch https://orcid.org/0000-0002-3510-5422

Notes

1. Of course, even if the stakes associated with a decision do not engender a right to explanation, there might be other reasons unrelated to stakes that do engender such a right.
2. The right to explanation may require justifications or motivating reasons-explanations as well. For instance, if a public servant makes a final decision in part based on an algorithmic input, the right might require that she states her motivating and/or justifying reasons for the decision.
3. Proponents of pragmatic/moral encroachment in epistemology claim that practical stakes affect the level of justification needed to count as knowing a proposition. Inspired by this, one might coin something like "encroachment on explanation," which suggests that a more capacious explanation is needed as the stakes increase (Bolinger 2020). We think this idea is plausible—higher stakes require better explanations—but we doubt the flipside of this idea, namely that "low stakes" would engender an entitlement to a worse as opposed to no explanation. In our optics, "knowing" and "being entitled to an explanation" are asymmetrical in an important sense: there are internal, lower-boundary epistemic standards to "knowing" such that you can only count as knowing if you have *some* evidence even in low stakes cases. But there are no comparable internal standards to "*being entitled to* an explanation" (even if there are internal standards to what it means to successfully explain something or understand something). If stakes matter here, and if the stakes are not sufficient to trigger the underlying concern that grounds the importance of stakes (e.g., autonomy), then no right to explanation is triggered.
4. A public sector parallel to Navigation may involve a local government, which uses an algorithm to plan traffic flow by letting it decide how long it takes for traffic lights to turn green. Over time, this may affect the travel time of daily commuters significantly.
5. For instance, as suggested by an anonymous referee, we can imagine chaining algorithms together: the output of one or more low-stakes algorithms can serve as input to another algorithm, whose output could then have high-stakes consequences.

References

Baron S (2023) Explainable AI and causal understanding: Counterfactual approaches considered. *Minds & Machines* 33: 347–377.

Beckers S (2022) Causal explanations and XAI. *Proceedings of Machine Learning Research* 140: 1–20.

Beigang F (2022) On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds & Machines* 32: 655–682.

Bhargava D and Gupta LK (2022) Explainable AI in neural networks using Shapley values. In: Khamparia A, Gupta D, Khanna A and Balas VE (eds) *Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI)*. Intelligent Systems Reference Library. Singapore: Springer, 59–72.

Bolinger RJ (2020) Varieties of moral encroachment. *Philosophical Perspectives* 34(1): 5–26.

Buijsman S (2022) Defining explanation and explanatory depth in XAI. *Minds and Machines* 32: 563–584.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Cao X and Yousefzadeh R (2023) Extrapolation and AI transparency: Why machine learning models should reveal when they make decisions beyond their training. *Big Data & Society* 10(1): 1–5.

Chou Y-L, Moreira C, Bruza P, et al. (2022) Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms and applications. *Information Fusion* 81: 59–83.

Coyle D and Weller A (2020) "Explaining" machine learning reveals policy challenges. *Science* 368(6498): 1433–1434.

Crisp R (2001) Well-Being. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy*, Winter 2021 Edition. Stanford: The Metaphysics Research Lab. https://plato.stanford.edu/archives/win2021/entries/well-being/.

Dietz A (2023) Pattern-Based reasons and disaster. *Utilitas* 35(2): 131–141.

Divyat M, Tan C and Sharma A (2019) Preserving causal constraints in counterfactual explanations or machine learning classifiers. CausalML: Machine Learning and Causal Inference for Improved Decision Making Workshop, 33rd Conference on Neural PRocessing Systems (NeurIPS2019), https://arxiv.org/abs/1912.03277

Eoin D, Pakrashi A, Greene D, et al. (2022) Counterfactual explanations for misclassified images: How human and machine explanations differ. https://arxiv.org/abs/2212.08733.

Fleisher W (2022) Understanding, idealization, and explainable AI. *Episteme; Rivista Critica Di Storia Delle Scienze Mediche E Biologiche* 19(4): 534–560.

Glover J and Scott-Taggart MJ (1975) "It makes no difference whether or not I do it.". *Proceedings of the Aristotelian Society, Supplementary Volumes* 49: 171–209.

Greta W, Keane MT and Byrne RMJ (2022) Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. https://arxiv.org/abs/2204.10152

Günther M and Kasirzadeh A (2022) Algorithmic and human decision making: For a double standard of transparency. *AI and Society* 37(1): 375–381.

Keane MT, Kenny EM, Delaney E, et al. (2021) If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), https://arxiv.org/abs/2103.01035

Kempt H, Freyer N and Nagel SK (2022) Justice and the normative standards of explainability in healthcare. *Philosophy & Technology* 4: 35–100.

Lazar S (forthcoming) Legitimacy, authority, and the political value of explanations. In: *Oxford Studies in Political Philosophy*. Oxford: Oxford University Press.

Lipton P (2001) What good is an explanation? In: Hon G and Rakover S (eds) *Explanation*. Dordrecht: Springer Verlag, 43–59.

Nefsky J (2015) Fairness, participation, and the real problem of collective harm. *Oxford Studies in Normative Ethics* 5: 245–271.

Parfit D (1984) *Reasons and Persons*. Oxford, GB: Oxford University Press.

Patty JW and Penn EM (2023) Algorithmic Fairness and Statistical Discrimination. *Philosophy Compass* 18(1): 1–23.

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

Purves D and Davis J (2022) Public trust, institutional legitimacy, and the use of algorithms in criminal justice. *Public Affairs Quarterly* 36(2): 136–162.

Robbins S (2019) A misdirected principle with a catch: Explicability for AI. *Minds and Machines* 29(4): 495–514.

Ross LN (2023) The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese* 202: 56.

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

Sahil V, Dickerson J and Hines K (2020) Counterfactual explanations for machine learning: A review. https://arxiv.org/abs/2010.10596

Selbst AD and Powles J (2017) Meaningful information and the right to explanation. *International Data Privacy Law* 7(4): 233–242.

Stefánsson HO (2023) The tragedy of the risk averse. *Erkenn* 88: 351–364.

Taylor E (2023) Explanation and the right to explanation. *Journal of the American Philosophical Association* forthcoming: 1–16. doi:10.1017/apa.2023.7.

Thoma J (2019) Risk aversion and the long run. *Ethics* 129(2): 230–253.

von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology* 34(4): 1607–1622.

Vredenburgh K (2022) The right to explanation. *Journal of Political Philosophy* 30(2 (June 2022)): 209–229.

Wachter S, Mittelstadt B and 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.

Wenar L (2023) The value of unity. *Philosophy & Public Affairs* 51: 195–233.

Wong YN, Jones R, Das R, et al. (2023) Conditional trust: Citizens' council on data-driven media personalisation and public expectations of transparency and accountability. *Big Data & Society* 10(2): 1–13.

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

Woodward J (2019) Some varieties of non-causal explanation. In: Reutlinger A and Saatsi J (eds) *Explanation beyond causation: Philosophical perspectives on non-causal explanation*. Oxford: Oxford University Press, 117–140.

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