Explanation and the Right to Explanation.

Elanor Taylor

Department of Philosophy, Johns Hopkins University

etaylo42@jh.edu

(Penultimate draft; please cite published version.)

Abstract

In response to widespread use of automated decision-making technology, some have considered a *right to explanation*. In this paper I draw on insights from philosophical work on explanation to present a series of challenges to this idea, showing that the normative motivations for access to such explanations ask for something difficult, if not impossible, to extract from automated systems. I consider an alternative, outcomes-focused approach to the normative evaluation of automated decision-making, and recommend it as a way to pursue the goods originally associated with explainability.

Keywords: Artificial Intelligence, Explanation, Right to Explanation

1. Automated Decision-Making and the Right to Explanation

An enormous number of decisions about our lives are made by automated systems, including decisions about mortgages, credit lines, healthcare, entertainment recommendations, policing,

sentencing, and personal matchmaking. Automated decision-making is more readily accessible than ever before, but its widespread use has raised public concern. For example, automated decision-making guides central parts of the criminal justice system in the United States, including parole decisions, and many have argued that this practice entrenches established patterns of judicial injustice (Angwin, Larson, Mattu & Kirchner 2016; Hao 2019; Metz & Satariano 2020). In 2020 automated grade calculations replaced in-person final examinations for many UK school students during the Covid pandemic. The results met with public outcry as calculations took expected performance from schools into account, resulting in lower grades for students from underperforming schools (Ascher-Shapiro 2020; Broussard 2020; Hern 2020; Zimmerman 2020). Even apparently benign applications of automated decision-making have come under scrutiny. For example, the algorithm that governs recommendations on YouTube has been implicated in an increase in political polarization and the spread of conspiracy theories, such as Covid denial and Q-Anon (Ovide 2020; NY Times 2020).

Automated decision-making systems are often *opaque*, such that the bases of their decisions are inaccessible to the average person. Opacity comes in different forms. Some systems are opaque simply in that their decision-making is kept secret by the organizations that create and use them. Sometimes opacity is more extreme, as in cases of *algorithmic opacity* where the system's decision-making procedures are unavailable even to its designers, because its decision-making capacity emerges out of interactions between its code and complex training data (Paudyal & Wong 2018:193). Such systems effectively train themselves to deliver results, and typically even their designers do not, and cannot, understand how they operate.

The widespread use and opacity of automated decision-making is worrying and has led some to consider a *right to explanation*. On this line of thought, if an automated system makes a decision about us, then we have a right to an explanation of that decision. Some take the European Union's 2018 General Data Protection Regulation (GDPR) to enshrine a right to explanation, though this is a matter of some legal controversy (Selbst & Powles 2017; Wachter, Mittselstadt & Floridi 2017a). According to Article 13 of the GDPR if a decision is made about a person through an automated process then that person should have access to "meaningful information about the logic involved" in the decision.[1] The GDPR applies only in the EU but similar legislation has been pursued elsewhere, such as New York City's LL 144 which regulates automated decision-making in hiring and human resources (Hunton Andrews Kurth LLP 2023). Furthermore, some have argued that independent of explicit legal regulation there are good ethical, political, and social reasons to embrace a more transparent approach to automated decision-making, a central motivation for developments in *explainable AI* (Paudyal & Wong 2018; Wachter, Mittelstadt & Floridi 2017b).

There has been a flurry of discussion about the challenges raised by the prospect of explainable AI and a right to explanation. For example, some argue that the GDPR does not legally guarantee a right to explanation, while others hold that explainable AI is impossible to engineer.[2] However, at the heart of this issue is a philosophical question: what is the right to explanation a right *to*? What does it *take* to explain an automated decision? There is little consensus on this matter. GDPR Article 13 uses the language of "meaningful information", while some appeal to

[1] European Union Regulation, Article 13, 2f. The GDPR, like other EU laws, is divided into *articles* that constitute the legal requirements and *recitals* that offer contextualization and interpretation of the articles.
[2] See Wachter, S., Mittselstadt, B. & Floridi, F. (2017a). Representatives of Google Research referred to machine learning as "alchemy" in a speech given in 2017. See Rahimi, A. & Recht, B. (2017)

the idea that an explanation is an answer to a why-question (EU GDPR A13, 2f; Burrel 2016). But these answers are too broad to be of much use. For example, the fact that a decision was made by an algorithmically opaque system is meaningful information, and offers an answer (of a sort) to the question of why the decision was made. But this is clearly an inadequate explanation for most purposes, and so more insight is necessary. In this paper, I take up this issue.

The normative motivations for access to explanations offer a valuable source of insight. By reflecting on why access to explanation might be a good thing, we can clarify what an explanation must be such that it can satisfy those motivations. Accordingly, I begin with the normative case for explainable AI. However, as I will show, the normative motivations for explainable AI appear in many cases to demand explanations that are difficult, if not impossible, to obtain from automated decision-making systems, including explanations that give information about reasons. This threatens the claim that there is a right to explanation. I conclude by considering some responses. I argue that opacity undermines the kind of evaluation that involves seeking reasons and explanations for decisions, and discuss some broader implications of this insight for normative exchange and deliberation in opaque contexts. I then sketch an alternative approach to automated decision-making that focuses on outcomes rather than explanations, and recommend this as an alternative way to pursue the goods originally associated with explainable AI. Many organization are currently moving away from pursuing explainability towards an emphasis on outcomes, and this discussion can be understood as offering a philosophical basis for such strategies.[3]

[3] For example, see Microsoft's Guidelines on Impact Assessment (Microsoft 2022), NYC L44 which focuses in part on outcome-focused *audits* (Hunton Andrews Kurth LLP 2023), and the EU AI Act Proposal (EU 2021).

2. Normative Motivations for Explainable AI

Public conversation about explainable AI is dominated by the language of rights, in particular the phrase "right to explanation." But showing that there is a right to explanation is different from showing that there is a normative case for access to explanations. The case for a right to φ must show that the normative case for φ is strong enough to justify the burden of providing φ. If that burden is considerable, the case for the right may be undermined.[4] I will exploit this distinction between making a normative case and establishing a right, and so will treat the case for explainable AI and the case for a right to explanation as distinct.

Let us begin by considering normative motivations for explainable AI.

One central motivation is to promote *transparency.* This is mentioned in the GDPR recitals:

> It should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. (EU GDPR, Recital 39)

Recently some commentators have explored this motivation. For example, Seth Lazar argues that automated decision-making intensifies existing political power relations and as such is subject to a publicity requirement, that it should be possible for those who authorize the power's use to

[4] I follow Kate Vredenburgh in adopting this framing of the right to explanation (Vredenburgh 2021). In doing so I treat the right to explanation as a positive rather than a negative right, and leave open questions about which more fundamental rights it may derive from.

determine that it is being used legitimately and with proper authority (Lazar ms: 8). Lazar argues that explainability is necessary to meet this publicity requirement, and so that it motivates the demand for explanation. Mario Günther and Atoosa Kasirzadeh go further, arguing that automated systems should be held to an even higher standard of transparency than human beings (Günther & Kasirzadeh 2022).

However, there is more to this issue beyond transparency or publicity, as there are further goods that transparency can promote. The GDPR recitals mention that access to an explanation of an automated decision will help us to evaluate that decision, and the decision-making procedure, for *fairness* (EU GDPR, Recital 39). This requires that the explanation not only provide information about the decision in an individual case, but also about how equitably the decision-making criteria are applied across populations. For example, all things considered, applying a higher salary requirement to one candidate over another for the same credit decision is a violation of fairness. This issue is central to public outcry about automated decision-making technology, much of which appeared in response to apparently unfair applications of decision-making, such as automated exam grade calculations and parole decisions.

A further good that transparency promotes is the protection of individual *autonomy* with respect to decisions made by automated systems. Access to an explanation can help us to protect our autonomy because it permits us to understand the bases of the automated decisions and, where appropriate and possible, exert control over our circumstances to achieve different results. If my mortgage application is refused, for example, it is reasonable to expect an explanation that tells me what I can change to raise the likelihood that a future application will be accepted. It is not always possible to change the outcome of a decision, as when a health insurance request is

rejected because of a pre-existing condition, but my access to that information still supports my autonomy by enabling me to understand that fact and base further actions upon it, such as seeking different health insurance or voting for changes to the healthcare system. Similarly if a decision is made through a random process, as in a lottery, then it is clear that I can take no further steps to change the outcome. Much discussion of the right to explanation has focused on the significance of autonomy (Vredenburgh 2021; Wachter, Mittselstadt, & Russell 2018). Some may disagree that protecting the pursuit of certain goals, as in these cases, is properly called "protecting autonomy." If so, we can adopt different language and say instead that access to explanations may be valuable in part because it facilitates goal-seeking.

Access to an explanation may be valuable because it allows us to evaluate the use of decision-making systems for its capacity to generate broader *social harm*. Such harm, which goes beyond instances of individual unfairness or violations of individual autonomy, has been the subject of much of the public conversation about automated decision-making. For example, as mentioned earlier, some commentators argue that the YouTube selection algorithm, which nudges viewers towards content similar to content they have already viewed, is implicated in online radicalization, the spread of conspiracy theories, and the undermining of public trust in mainstream news sources and scientific institutions (Ovide 2020; NY Times 2020). Facebook has been subject to similar scrutiny for news and content recommendations that trap users in "filter bubbles" and so hamper the exchange of ideas essential to the democratic process, and facilitate the spread of misinformation and conspiracy theories (Adee 2016; BBC 2021/2). From these examples we can see that social harm can often be unintentional and hard to predict, but transparency about the bases of the decisions made by automated systems will allow us to at least evaluate the use of such systems for the likelihood of generating social harm. A central

characteristic of social harm is that it is structural, affecting society at the level of systems and institutions, and as such is not equivalent to individual-level injustice, though it inevitably generates harm to individuals.

This list of motivations is not exhaustive but it offers some sense of why explainable AI might be a good thing. Access to explanations of automated decisions will promote transparency, which is valuable in turn because it allows us to evaluate those decisions for fairness, and/or to protect our autonomy with respect to the outcomes of the decisions, and/or to evaluate the use of the technology for its potential to cause social harm. These goals are general, in that widespread access to explanations will promote them in general, but not every attempt to explain a decision made by an automated system must in itself meet all of these demands. For instance, an explanation that permits an individual to protect their autonomy, such as an explanation that identifies the factors relevant to a credit decision, may not be rich enough to facilitate evaluation of the decision-making system for its capacity to generate social harm. This does not undermine the value of the explanation, as it still permits the individual to protect their autonomy and thereby satisfies one of the normative motivations for explainable AI. More broadly, however, even if each attempt at explanation need only meet the needs of those seeking it, in general people should have access to explanations that satisfy *all* of the motivations described here.

3. Normatively Satisfying Explanations

Having examined the normative motivations for explainable AI let us now consider what explanation is such that it can satisfy these motivations.

There is little consensus among philosophers about the nature of explanation. Philosophers disagree on almost every aspect of the subject including whether or not explanation must be causal, whether or not it must be factive, and what kind of relation it is, if it is a relation at all (Achinstein 1983; Craver 2014; Lange 2016; Lewis 1986; Reutlinger 2017; Ruben 1990; Salmon 1989; Skow 2014; Van Fraassen 1980). Indeed, some have argued that it is misguided to attempt to give a general account of explanation (Díez, Khalifa & Leuridan 2013; Nickel 2010). If we turn to philosophy for insight into the right to explanation then this level of disagreement might seem frustrating. But thankfully we do not need a full theory of explanation to illuminate this issue. Instead, I will focus on two claims about explanation that, while not universally endorsed, are fairly uncontroversial. The first is that explanation has a contextual aspect, and the second is that we can at least partially explain events by identifying difference-makers.

The idea that explanation has a contextual aspect is simple to motivate. Imagine that I ask you to explain a particular event: the theft of a car. You would naturally consider features of our context before offering the explanation. If we are engineers and are interested in the car theft as a physical event, then you would be likely to offer a physical, but not micro-physical, explanation of the theft, giving information about the forces exerted upon the car, how the glass was broken, and so on. Alternatively, if I am your friend and am asking what possessed you to steal this car, then a different sort of explanation will be appropriate. In this situation I want an explanation that tells me about your motivations, and what inspired you to steal the car. This contextual aspect of explanation is commonplace, and some accounts rest more significance on it than others. For example, some hold that what counts as an explanation is determined at least partly by context (Achinstein 1983; Van Fraassen 1980). On other views explanation itself is context-independent, but context affects how successfully the explanation may be received or understood

(Lewis 1986). We need not decide on this issue here, and may simply acknowledge that explanation has *some* contextual aspect, as follows:

> CONTEXT: Successful explanations are contextually appropriate.

The second claim, that we can explain events by identifying difference-makers, is more controversial though still widely-endorsed (Woodward 2003). The rough idea is that we can explain an event by identifying the factors that made a difference to the occurrence of the event, picked out by the counterfactual test: had those factors had not been in place, the event would not have occurred. For example, if it is the case that, had I not thrown the rock, the window would not have broken, then the information that I threw the rock will (at least partially) explain the event of the window breaking. The difference-making approach to explanation has this idea at its heart and captures a number of intuitively plausible claims about explanation such as that explanation has a counterfactual aspect, and that there is a close connection between explanation and intervention. The full difference-making approach to explanation is detailed, and includes a formalism for representing causal, and hence explanatory, relations in structural equation models (Pearl 2000; Woodward 2003). However, most of that detail is irrelevant to this conversation. All we need here is the rough idea that we can (at least partially) explain events by identifying difference-makers:

> DIFFERENCE: We can (at least partially) explain an event by identifying factors that made a difference to that event.

There are good reasons to adopt DIFFERENCE as a frame for this issue. First, many normative motivations for access to explanations of automated decisions appear to ask for information about difference-makers, such as the case in which the information that a lack of long-term loan

history made a difference to a credit denial satisfied the requirements of transparency and of protecting autonomy. Second, the output of an automated decision-making system is an event. Despite the lack of consensus about explanation in general, difference-making is the most commonly-accepted approach to explanation of events in philosophy of science, which makes this a reasonable place to start (Reutlinger 2017). Third, although the literature on what "explanation" means in "right to explanation" is in a developmental stage, the proposal that explanations must identify difference-makers has been the subject of some discussion (Wachter, Mittselstadt, & Russell 2018). Fourth, the proposal as it stands is not that *all* explanation, or all causal explanation, proceeds by identifying difference-makers, but is instead the weaker claim that one *can,* at least partially, explain events by identifying difference-makers. Accordingly, one need not endorse a full difference-making approach to explanation to endorse this proposal.

Some may worry that a difference-making approach is inapplicable to decisions made by automated systems because such explanations must be non-causal. I remain neutral on whether explanations of automated decisions are causal or non-causal, but note that although difference-making was developed as a model of causal explanation, it has been extended to non-causal explanation (Reutlinger 2018; Schaffer 2016; Wilson 2018. For challenges see Kasirzadeh 2020). Others may worry that difference-making explanations are impossible to extract from many automated decision-making systems, particularly black-box systems (Grimsley ms). I will return to this issue in Section 5, where I will suggest that if they are impossible to extract, then this reveals a tension at the heart of the explainable AI project.

Putting CONTEXT and DIFFERENCE together gives a clearer picture of the target of the normative motivations for explainable AI. To explain a decision made by an automated system

we should identify difference-makers – factors such that, had they not been in place, the decision would have gone differently – in a contextually appropriate way. The standards for contextual appropriateness are found in the normative motivations discussed in Section 2. These explanations must promote transparency and must enable us to evaluate the decisions for fairness, and/or permit the inquirer to protect their autonomy, and/or permit evaluation for the capacity to generate broader social harm, depending on the goals of the inquirer:

> TARGET: Normatively satisfactory explanations of automated decisions identify factors that made a difference to the decision in a manner that promotes transparency, permits evaluation of the decision and decision-making system for fairness and/or capacity to generate social harm, and/or enables the inquirer to protect their autonomy.

As before, the normative motivations for explainable AI do not require that every explanation should meet each of the individual criteria. Depending on the interests of the inquirer, in some cases an explanation that permits evaluation for fairness but is not rich enough to support evaluation for social harm will be sufficient. But overall, the normative case for explainable AI supports widespread access to explanations that meet all of these criteria. Let us now consider what it takes for an explanation to meet these standards.

One obvious criterion is that to play any of these roles an explanation must be understandable to its audience. This follows straightforwardly from the transparency requirement. A second criterion, also obvious, is that the explanation must give information about factors that made a difference to the decision. It may be tempting to stop here, holding that all that is required is to merely identify some factor that made a difference to the decision, and to do so in an understandable way. But this is the point at which the contextual aspect of explanation becomes

salient. In many contexts *merely* identifying *some* difference-maker will not satisfy the normative motivations for explainable AI. Instead, TARGET specifies that we need information about difference-makers that is particularly suited to the inquirer's goals.

For insight, consider the range of reasons why you may want an explanation of an automated decision. In some cases you may simply want to protect your own interests. When my credit application is denied, for example, I want to know what action I can take to improve the likelihood of a positive decision next time. For this I need information about a difference-maker pitched at a level that gives me control over future outcomes, where possible. The information that the system is designed to avoid credit risk and regarded my application as risky fails to provide this, while identifying a particular aspect of my case, such as my lack of loan history, is more successful. Sometimes, however, our needs are more complex, as when we are evaluating decisions and decision-making systems for their capacity to generate social harm. To perform these kinds of tasks we sometimes need information about a difference-maker pitched at a very particular level of grain.

Consider some examples. Imagine that my application for a mortgage is turned down and I am told that my lack of long-term loan history made a difference to the decision. This is a difference-maker, so if merely identifying a difference-maker is sufficient for explanation, then the task is complete. But I might reasonably wonder *why* long-term loan history made a difference. Perhaps it is important because it is independently motivated as an indicator of potential for responsible credit use, say. But alternatively, perhaps long-term loan history is important simply because it is correlated with affluence. These differences are significant when it comes to evaluating the decision-making for fairness and for social harm, as the latter basis for credit decisions is arguably

less fair than the former, and has potential to entrench existing inequality. This example indicates that sometimes we need to identify difference-makers at different levels of grain, where the appropriate level of grain is set by what the inquirer wants from the explanation.

Consider an alternative case. Imagine that I want to know why YouTube selected a particular video, outlining a conspiracy theory, for me to watch next. A difference-making explanation might tell me that this video is similar to those I have watched before, and so the difference-maker just is that this content is similar to content I have selected and watched in the past. The algorithm is designed to keep viewers watching, and similarity is a factor that will keep viewers watching. That seems fairly benign, though it might lead to some overconsumption, but if we keep asking questions a new picture may emerge. Imagine that the next video was selected because the content is similar. Furthermore, the content I watched previously is about conspiracy theories, and people who watch that kind of content are more likely than others to binge-watch YouTube videos. This is a finer-grained instance of the coarser difference-maker "similarity of content." These reasons are more exploitative, and this politically-relevant information is not available from the more general claim that this video is similar to content I watched before. In each case "similarity of content" is a difference-maker. But the reason why "similarity of content" is selected as a difference-maker is different in each case. In the first, it was selected simply because a viewer is more likely to watch similar content. In the second, it was selected because the viewer of a particular, politically-salient type of content is more likely to watch similar content. If we are interested in evaluating decisions and decision-making systems for fairness, autonomy, and social harm, such differences make a difference.

As we can see from these examples, the explanations needed to satisfy the normative motivations for explainable AI must get to a particular level of detail, set by the interests of the inquirer. And in some cases, the correct level of detail is for the explanation to identify a *reason*. A reason explanation is an explanation of an agent's action that gives information about the agent's own motivations for their action. If I explain why I donated to a particular charity by pointing out that I took it to be the morally right thing to do, then I am giving you a reason explanation for my action. And in many cases, to satisfy the normative case for explainable AI we need to get to the reasons behind a decision, rather than merely to identify any difference-maker.[5] For instance, in the mortgage application case we are interested in the reason why the loan history made a difference, and in the YouTube case we want to know the reason why similarity of content was selected as a difference maker. In each case we want the justification and motivation for that factor making a difference, not just that it *did happen* to make a difference. Without that information, we cannot meet our normative goals.

4. Revisiting the Right to Explanation

I have proposed that the normative case for explainable AI is a case for widespread access to contextually-appropriate difference-making explanations of the decisions made by automated systems. Contextual appropriateness varies from case to case, depending on the inquirer's interests, and the factors that determine contextual appropriateness include the capacity of the explanation to promote transparency, evaluation for fairness, individual autonomy, and evaluation for social harm. To meet these standards in some cases the explanation must provide

[5] On a standard view of reasons as causes, reasons are difference-makers. However, the view of reasons as causes has been challenged. See discussion in Queloz, M. (2018) and Sehon, S. (2000)

reasons for the decisions, and reasons why certain other difference-makers were selected. Let us now return to consider the implications of this proposal for the claim that there is a right to explanation.

As mentioned in Section 2, establishing a right to φ is different from making a normative case for φ, as the former must show that the normative case for φ outweighs or justifies the burden of providing φ. This exploration of normative motivations for explainable AI and the kinds of explanations that meet them reveals serious challenges to the claim that there is a right to explanation.

The first problem is simple: it is not clear that difference-making explanations of the decisions made by automated systems are reliably available. In simple cases, as when systems operate as decision trees, or operate with a small number of predictable parameters, there may be no in-principle barrier to identifying difference-makers. But in more complex cases opacity may arise from a range of sources, such as the complexity and high-dimensionality of the decision-making systems, and the inscrutability of the mathematical methods the systems use to develop decision-making procedures (Lazar ms:2-4). Each source of opacity generates different normative and epistemic challenges, and each presents barriers to widespread access to difference-making explanations. The nature and depth of the opacity of automated decision-making systems is a lively topic in computer science, so it is an open empirical question, and a matter of some controversy, whether viable difference-making explanations of the decisions made by automated systems are or will soon be available (Grimsley ms; Hamon, Junklewitz, Malgieri, Hert, Beslay, & Sanchez 2021). But in so far as features of the systems themselves, or of the organizations that use

them, preclude access to difference-making explanations, the normative motivations for explainable AI cannot be met.

In response one might consider an idealized standard for explainability, along the following lines: that some explanation for the decision must be available in principle, even if no actual person can currently access it. This standard would avoid problems of intractability. However, the normative motivations for access to explanations are based around the idea that it is a social good for *actual people* to have access to these explanations for their own purposes, not that an idealized agent could in principle access the explanation. Even though an idealized standard for explanation may avoid this challenge, it also fails to meet the normative motivations for explainable AI.

Second, in some cases the normative motivations for explainable AI are best satisfied by a reason explanation, but it is not clear that reason explanations of automated decisions are available, or that the prospect of a reason explanation of a decision made by an automated system is even coherent. This is partly because of the engineering problems involved in identifying difference-makers in opaque systems, as above. But it is also about the very idea of asking for reasons from an automated system. This is a strange demand, and it leaves us searching for candidate sources of reasons in the decisions of the engineer, of the organization conducting the decision-making, or even society. It is natural to find such a request jarring, as reason explanation is the kind of explanation that we use to explain specifically human affairs. Mechanistic and other kinds of causal explanation may be satisfactory when it comes to events in natural and artificial worlds, but when we want to understand human action and decision, we turn to reason explanation. These are the explanations that we ask for when we want to understand one another, and to hold

one another accountable. In so far as the normative case for explainable AI asks for reason explanations, it asks us to explain how automated decisions are made in a specifically human way.

Third, putting aside these concerns about opacity and the apparent absence of reasons, the resulting picture of the right to explanation is radical. The case for access to explanations of automated decision-making requires a significant level of information about corporate and institutional decision-making. If we take that case seriously, then it seems that we should demand a level of scrutiny of institutions using automated decision-making far beyond what was ever applied to actuaries, bank-managers and the like, regardless of the automation or otherwise of their decisions. To countenance a right to explanation is therefore to countenance an overhaul of industries and services including banking, computing, social media, education, medicine, government, and policing. If the normative case for access to such explanations is significant then this overhaul is required, and the fact that it may be challenging to implement is not in itself an objection. But it does put an extra burden on the positive case for a right to explanation, which must be very strong to overcome these practical challenges.

Each consideration generates a problem for the case for a right to explanation. The first two may show that a right to explanation is literally impossible, because in many cases the case for explainable AI appears to ask for something that cannot be provided. Putting aside those worries we are left facing the third concern, that providing such explanations demands an unprecedented level of transparency from any industry or organization using automated decision-making. This presents a significant burden, making it harder to show that the normative case for access to such explanations outweighs the cost of providing them. It may be, of course, that the normative case

for explainable AI meets this standard, or that a high level of transparency is motivated by other considerations. But overall, these are serious challenges to the claim that there is a right to explanation.

5. Responses and Reflections: Opacity and the Shift to Outcomes

One response to these challenges is to return to my reasoning through the normative motivations for explainable AI and reject some of that framing, arguing that an alternative picture offers a more promising outlook for a right to explanation. Perhaps a non-difference-making form of explanation is available from algorithmically opaque systems, or perhaps information about reasons is never required to meet the normative motivations for explainable AI. I will not consider such strategies here as I suspect that most explanations that can meet the standard of TARGET will generate similar problems, but there is space to attempt to show otherwise. A more pessimistic alternative is to abandon the normative evaluation of automated decision-making, and with it the idea that there is a right to explanation. On this line of thought the normative motivations for explainable AI ask for something that can never be delivered, so attempts to normatively evaluate automated decision-making are incoherent.

An in-between strategy involves seeking explanations of some automated decisions but not others. For example, Alex London points out that many clinical recommendations with a strong evidence base are not well-explained, even without the influence of automated decision-making (London 2019). Because opacity is a standard feature of medical decision-making, London argues, it is misguided to demand explanations of all automated clinical decisions. To take this line of reasoning is not to give up on normative evaluation of automated decision-making, but

instead to acknowledge that some contexts have epistemic features that can preclude explainability, which undermines a general demand for explanation.

While each of these strategies has its merits, here I will consider moving away from explainability for different reasons. London is correct that it is misguided to seek explanations of decision-making in opaque contexts, and the reasons why reflect deeper philosophical considerations about what it is to explain a decision. Seeking explanations of decisions in order to normatively evaluate them is to engage with those decisions in a broadly reason-giving way. In opaque contexts, however, the standard bases for reason-giving evaluation are often unavailable. A viable alternative is to switch the focus of evaluation to the outcomes of decisions, rather than the reasons for them. In the case of automated decision-making this entails a switch from evaluating decisions by explaining them to evaluating decisions by examining their outcomes. In the remainder of this paper I will sketch this idea in more detail. Given that many organizations and institutions are abandoning explainability in favour of the evaluation of outcomes for expediency, this discussion can be understood as an attempt to offer a more robust normative basis for this shift, and to recommend it as an alternative way to pursue the goods originally associated with explainable AI.[6]

To seek an explanation of a decision, especially an explanation that articulates reasons, is to treat the source of the decision as an entity like the agent seeking the explanation: as a rational being. In asking for these explanations we seek reasons for, justifications of, and motivations for the decisions, and thereby focus our evaluation on the rationale for the decision, rather than its

[6] For example, see Microsoft's guidelines on impact assessment (Microsoft 2022), NYC L44 which focuses in part on outcome-focused *audits* (Hunton Andrews Kurth LLP 2023), and the EU AI Act Proposal (EU 2021).

outcomes. As such, this is a traditionally deontological mode of normative evaluation (though some approaches to deontology are more reason-focused than others). However, in the context of automated decision-making the standard sources of such evaluation, like sentience or rationality, are not in play. The systems we seek these explanations from do not have the features typically taken to ground reason-giving interactions and evaluations. In the same way that it is wrongheaded to blame a tree for dropping a branch onto my car, or to ask the cloud to explain why it rained on me, it is wrongheaded to ask an automated system to explain its actions, and especially to give an account of its reasons.

Automated decision-making systems may have these features by proxy. While an algorithm in itself may not have reasons, its decision-making may reflect the reasons of the company or institution using it. However, as we have seen, when automated decision-making is used, often *those* reasons are *also* kept opaque. The automated system may display its own kind of opacity, and the company or institution using the system often imposes further layers of opacity. Even if we take the automated system to have reasons by proxy, we typically still do not have access to those reasons.

The normative case for explainable AI is based on the idea that access to explanations of automated decisions will promote transparency, and help us to ensure that we are treated fairly, can protect our autonomy, and can protect society from broader social harms. As we have seen, considerations about opacity, intractability, and the absence of reasons threaten to undermine this focus on explanation, and warrant turning towards a more tractable place: the *outcomes* of automated decision-making. This shift of focus is in the spirit of a political philosophy which recognizes that harms are easily generated by well-intentioned systems, and takes unjust

outcomes as sufficient to warrant political response, without requiring evidence of unjust intent.[7] If automated decision-making consistently, say, privileges the rich, promotes racist judicial policies, hampers social mobility, and breaks down information channels essential to democratic communication, then, on this line of thought, *that is enough to be working with*, without having to also scrutinize an opaque system for the source of the decisions that generated such harms.

Consider a comparison with similar reasoning in a very different context: feminist theories of misogyny. Traditionally misogyny was understood in psychological terms, such that the misogynist, by definition, hates women. In contrast, Kate Manne recommends understanding misogyny as a social phenomenon. On this approach misogyny is something that women experience and face, rather than something that misogynists think (Manne 2017: 59-67). In adopting this view Manne recommends a shift from a mode of evaluation and engagement that focuses on the misogynist's reasons for their actions, to a mode of evaluation that focuses on the outcomes of those actions in women's lives.

The details of Manne's view and its place in broader discourse about misogyny are not relevant to this discussion. But the structure of this shift, from regarding the misogynist as a source of reasons to merely a source of outcomes, mirrors the shift I recommend in the case of automated decision-making. This is a move from considering the *reasons why* automated decisions are made to considering *what the decisions do*. This parallel is not coincidental, because both contexts display forms of opacity that undermine reason-giving evaluation. In political discourse under gender inequality (and similar conditions such as racism), the norms of reason-giving, explaining,

[7] Recent work on algorithmic bias adopts a similar strategy, such as Johnson, G. (forthcoming), (2021)

justifying, and so on, can often break down. The person who shifts their focus from reasons to outcomes in such situations treats the source of harm as opaque.

Structural harm such as misogyny is not required to justify this change in mode of evaluation, because other sources of opacity suffice. Different sorts and sources of opacity may generate different results here. For example, consider the opacity of decision-making in a company that keeps its procedures secret. One might reasonably adopt a consequences-forward approach to their decisions on the basis of the belief that, so long as the bases of their decisions are unavailable, it makes no sense to recognize them as a proper subject of reason-giving evaluation. If the company were to change policy in favor of a more transparent approach, one could then reasonably ask for explanations, and for information about their reasons and motivations. Alternatively, the kind of opacity evident in algorithmically opaque systems makes the search for explanations less tractable, and the mere fact that the decision-maker is an automated system undermines treating it as an agent with reasons in the way that full deontological evaluation (of this reason-focused sort) demands.

These considerations indicate that in opaque contexts, where the bases of a decision-maker's decisions and their decision-making procedures are unavailable for scrutiny, it is reasonable, and perhaps necessary, to adopt an outcomes-based model of normative evaluation. Accordingly, for those who find the normative case for explainable AI compelling and are worried about opacity and intractability, a shift in focus to the outcomes of automated decision-making offers a way to pursue the goods originally associated with access to explanations.

Precisely how well a focus on outcomes can serve the various normative motivations for explainable AI is unclear, and a developed proposal will have to show how information about outcomes can promote fairness, protection of individual autonomy, and so on. But overall, abandoning the focus on explanation as the core of normative concerns about automated decision-making and shifting instead towards evaluating outcomes offers a way to pursue the benefits originally associated with access to explanations, without the associated problems of intractability and implementation.[8]

[8] Thanks to Ramón Alvarado, Michael Brent, Nina Emery, James Goodrich, Neil Manson, Elizabeth Miller, Dean Moyar, Meghan Page, Kyle Rawlins, Erica Shumener, John Symons, and Michael Titelbaum, for helpful discussion and feedback. Particular thanks to Tal Linzen for an introduction to the topic, and for discussion and comments on an early draft. Thanks to audiences at "Computational Methods and the Future of Science" at the University of Kansas, a philosophy department colloquium at the University of Mississippi, and the 2021 meeting of the Philosophy of Science Association. Thanks to two online groups for workshopping the paper: the WiM Research Network, and the Pre-Tenure Women's online working group. Finally, thanks to two anonymous referees from this journal for their generous feedback.

References

Achinstein, P. (1983) *The Nature of Explanation.* Oxford University Press.

Adee, S. (2016) "How can Facebook and its users burst the 'filter bubble'?" *New Scientist* https://www.newscientist.com/article/2113246-how-can-facebook-and-its-users-burst-the-filter-bubble/

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) "Machine Bias." *Propublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ascher-Shapiro, A. (2020) "Exam grading algorithms amid coronavirus: what's the row about?" *Reuters* https://www.reuters.com/article/global-tech-education-idUSL8N2FD0EI

Broussard, M. (2020) "When Algorithms Give Real Students Imaginary Grades." *New York Times* https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html

Burrel, J. (2016) "How The Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 1-12

Craver, C. (2014) "The Ontic Account of Scientific Explanation." In Kaiser, M, Scholz, O.R., Plenge, D. & Hüttemann, A. (2014) (eds.) *Explanation in the Special Sciences: The Case of Biology and History.* Springer Verlag. pp. 27-52

Díez, J., Khalifa, K., Leuridan, B. (2013) "General Theories of Explanation: Buyer Beware." *Synthese* 190 pg 379-396

European Union (2021) Proposal: *The Artificial Intelligence Act* https://artificialintelligenceact.eu/

European Union, Regulation (EU) General Data Protection Regulation. Accessed at https://gdpr-info.eu

Grimsley, C. (manuscript) "Causal and Non-Causal Explanations of Artificial Intelligence."

Günther, M. & Kasirzadeh, A. (2022) "Algorithmic and human decision making: for a double standard of transparency." *AI and Society* 37(1):375-381

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P., Beslay, L., Sanchez, I. (2021) "Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario." *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* https://doi.org/10.1145/3442188.3445917

Hao, K. (2019) "AI is sending people to jail—and getting it wrong." *MIT Technology Review* https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Hern, A. (2020) "Ofqual's A-level algorithm: why did it fail to make the grade?" *The Guardian* https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels

Hunton Andrews Kurth LLP (2023) "NYC DCWP Proposes Rules to Implement New Law Governing Automated Employment Decision Tools." *National Law Review* Vol 12 No 304 https://www.natlawreview.com/article/nyc-dcwp-proposes-rules-to-implement-new-law-governing-automated-employment-decision

Johnson, G. (2021) "Algorithmic bias: on the implicit biases of social technology." *Synthese* 198:10 9941-9961

Johnson, G. (forthcoming) "Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning." *Journal of Moral Philosophy*

Kasirzadeh, A. (2020) "Counter Countermathematical Explanations." *Erkenntnis* https://doi.org/10.1007/s10670-021-00466-x

Lange, M. (2016) *Because Without Cause.* Oxford University Press.

Lazar, S. (manuscript) "Legitimacy, Authority, and the Political Value of Explanations." Accessed from Philpapers archive: https://philpapers.org/archive/LAZLAA-2.pdf

Lewis, D. (1986) "Causal Explanation." In *Philosophical Papers Volume II*. Oxford University Press.

London, A. (2019) "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49(1):15-21

Manne, K. (2017) *Down Girl: The Logic of Misogyny.* Oxford University Press.

Metz, C & Satariano, A. (2020) "An Algorithm That Grants Freedom, or Takes it Away." *New York Times* https://www.nytimes.com/2020/02/06/technology/predictive-algorithms-crime.html

Microsoft (2022) "Responsible AI Impact Assessment Guide." https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf

Nickel, B. (2010) "How General Do Theories Of Explanation Need To Be?" *Nous*. 44(2) 305-328

Ovide, S. (2020) "Take YouTube's Dangers Seriously." *New York Times* https://www.nytimes.com/2020/04/20/technology/youtube-conspiracy-theories.html

Paudyal, P. & Wong, W. (2018) "Algorithmic Opacity: Making Algorithmic Processes Transparent through Abstraction Hierarchy." *Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting.*

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Queloz, M. (2018) "Davidsonian Causalism and Wittgensteinian Anti-Causalism: A Rapprochement." *Ergo* 5(6):153-72

Reutlinger, A. (2017) "Explanation Beyond Causation? New Directions in the Philosophy of Scientific Explanation." *Philosophy Compass* DOI 10.1111/phc3.12395

Reutlinger, A. (2018) "Extending the Counterfactual Theory of Explanation." In Reutlinger, A., and Saatsi, J. (eds.) *Explanation Beyond Causation. Philosophical Perspectives on Non-Causal Explanations*. Oxford University Press.

Ruben, D-H. (1990) *Explaining Explanation.* Routledge.

Salmon, W. (1989) *Four Decades of Scientific Explanation*. University of Pittsburgh Press.

Schaffer, J. (2016) "Grounding in the Image of Causation." *Philosophical Studies* 173 (1):49-100

Sehon, S. (2000) "An Argument against the Causal Theory of Action Explanation." *Philosophy and Phenomenological Research*, 60 (1): 67-85

Selbst, A.D. & Powles, J. (2017) "Meaningful Information and the Right to Explanation." *International Data Privacy Law* 7(4): 233–242

Skow, B. (2014) "Are There Non-Causal Explanations (of Particular Events)?" *British Journal for the Philosophy of Science* 65 (3): 445-467

Van Fraassen, B. (1980) *The Scientific Image.* Oxford University Press.

Vredenburgh, K. (2021) "The Right to Explanation." *The Journal of Political Philosophy*. https://doi.org/10.1111/jopp.12262

Wachter, S., Mittselstadt, B. & Floridi, F. (2017a) "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. & Floridi, L. (2017b) "Transparent, Explainable, and Accountable AI for Robotics." *Science (Robotics)* 2 (6):eaan6080

Wachter, S., Mittselstadt, B. & Russell, C. (2018) "Counterfactual Explanations Without Opening The Black Box: Automated Decisions And The GDPR." **arXiv:1711.00399 [cs.AI]**

Wilson, A. (2018) "Metaphysical Causation." *Noûs* 52 (4):723-751

Woodward, J. (2003) "Making Things Happen." Oxford University Press.

Zimmerman, A. (2020) "The A-level results injustice shows why algorithms are never neutral." *New Statesman* https://www.newstatesman.com/politics/2020/08/the-a-level-results-injustice-shows-why-algorithms-are-never-neutral

Other Media

NY Times podcast (2020) "Rabbit Hole." https://www.nytimes.com/column/rabbit-hole

BBC podcast (2021/22) "The Coming Storm." https://www.bbc.co.uk/programmes/m001324r

Rahimi, A. & Recht, B. (2017) Award Speech from NIPS 2017. Transcript: http://www.argmin.net/2017/12/05/kitchen-sinks/