Undergraduate Philosophy Journal of Australasia www.upja.online
VOL. 4, DECEMBER 2022

AI and the Value of Explanations

BENJAMIN ROBINSON[1]

AUSTRALIAN NATIONAL UNIVERSITY

Abstract

AI systems often struggle to explain their outputs. Some have argued that this lack of explainability justifies banning their use in certain contexts. However, this paper argues that in many cases where AI systems are being deployed, the types of explanations we want from these tools can be provided by current methods in explainable AI. I make this case by first distinguishing between the instrumental and non-instrumental reasons why explanations are important. I then apply this analysis to the types of explanations we're able to obtain from AI systems using current explainability methods. This aims to demonstrate that the general, correlative - though not causal - reasons that explainable AI techniques provide are often sufficient for our interactions with large institutions, like governments, hospitals and banks, where explanations are instrumentally important in helping individuals understand, challenge and improve decisions. There are some cases, however, where explanations are non-instrumentally important in that they evidence respect for people as such (end of life care), or where specific causal reasons matter (criminal sentencing), which AI cannot provide. I discuss objections throughout, and finish with the caveat that while explainable AI methods are available in

[1] Ben Robinson is a student in Philosophy at the Australian National University. His work primarily revolves around AI ethics and value.

theory, the time and cost it takes to properly implement these techniques means that regulation or other forms of incentives are likely needed to ensure they are actually used in practice.

1. **Introduction**

AI systems are being increasingly used throughout society; from accessing finance (Booth 2019), to criminal sentencing (Kleinberg et al. 2017), to diagnosing cancer (Yala et al. 2021), to executing drone strikes (Lee 2021). Trained on huge datasets and computing power, these systems are predictively powerful, but we often don't know how or why decisions were made. Unlike traditional algorithms, where rules are pre-specified by human engineers, modern AI systems using machine learning algorithms essentially create their own rules to improve their performance (Mittelstadt et al. 2016), meaning ex ante predictions or ex post assessments of the system's operations are difficult to obtain (Zerilli et. al. 2019).[2]

There are a growing number of examples evidencing the harms of opaque automated decision making,[3] as well as regulatory proposals to enshrine some right to explanation in law.[4] Researchers like Vrendenburgh (2021) have given an account for an individual right to explanation focussing on the rights of decision subjects, while Lazar (2022) gives an account of the duties owed to the political community at large. Some argue that the demand for explainability is overblown; human decision making is just as, if not more, inscrutable than AI because of bias and motivated reasoning (Zerilli et al. 2019). Others argue that we should avoid using AI all together until AI systems can give "full and satisfactory explanations" for the decisions they make (House of Lords, 2019).

While there is something intuitively plausible about the explainability objection, I argue that strong forms of this objection misunderstand both the types of explanations needed

[2] Throughout the essay, I refer interchangeably between Artificial Intelligence (AI), AI systems, and automated decision making; and between explainability, inscrutability and opacity. Roughly, AI is any type of computational system that shows intelligent behaviour conducive to reaching goals (Muller 2021). Explainability is the capacity for AI systems to generate intelligible reasons for their outputs (McDermid et al. 2021). Section three of this essay goes into greater detail about both sets of definitions.

[3] In Australia, it was reported that more than 2000 people died after receiving an incorrect government debt notice without adequate explanation or right of reply (Medhora 2019).

[4] For example, Article 15 of the EU's GDPR aims to enshrine some form of explanation in law.

in different contexts where AI systems are being deployed, as well as the forms of explanations possible from modern AI systems. In what follows, I aim to demonstrate that methods in explainable AI that give general correlative, though not causal reasons for why a decision was made, are often sufficient for our interactions with large institutions like governments, hospitals and banks, where explanations primarily allow for understanding, challenging and improving decisions. However, there are cases, like in criminal sentencing or end-of-life care, where specific causal reasons matter, or where explanations evidence respect for people as such, and AI is unable to provide such explanations. So, overall, the demand for explainability can often, though not always, be met.

To make this case, I first distinguish between different instrumental and non-instrumental reasons why explanations are valuable (section two). I then analyse why AI systems are said to be opaque, and methods to limit this opacity (section three). Section four brings these strands together, where I argue that the instrumental value of explanations - challenging and improving decisions - are generally what's required for decision subjects interacting with large institutions, and these can be provided by techniques in explainable AI. Section five provides a contrary view, arguing that sometimes explanations tracking non-instrumental value - instantiating respect and acting for the right reasons - will be needed, and these cannot be given by AI. I discuss objections as I go, and I finish by showing how the analysis approach outlined in this essay can be used to make sense of individual cases where there is a demand for AI explainability.

2. **Why do explanations matter?**

I begin by outlining the instrumental and non-instrumental value of explanations, before applying this analysis to AI.

AI and the Value of Explanations

To explain something is to communicate information that enables an audience to reach a justified understanding of it (Wilkenfeld 2014).[5] The mere feeling of understanding isn't enough; it needs to be justified, which could involve explicating reasons, beliefs, intentions, external causes (Malle 2004) or deliberative and institutional procedures that influence a decision (Doshi-Velez et al. 2017). What evidence is drawn upon will be relative to an audience's goals; explanations enabling a data scientist to improve an AI system will be different from explanations enabling prediction recipients to understand why their loan was declined. The former will require more technical details enabling a justified understanding of how to improve the internal workings of the system; the latter requires information about how loan decisions are generally made, including factors like savings, debt and credit history. So, explanations serve different purposes in different contexts.

There are at least two broad reasons why explanations are instrumentally valuable. First, they allow decision subjects to challenge decisions and engage in "informed self-advocacy" (Crawford and Schultz 2014; Vredenburgh 2021). If we're told by a government decision making system that we owe the tax or welfare department money, we will want to know how this decision was made so we can challenge it if we think it's wrong. Challenging also applies to other stakeholder groups, like regulators, who require banks to explain how they determine credit scores to ensure compliance with discrimination law. This is linked to accountability, which generally requires that we are able to track who has made certain decisions. It can also identify underlying reasons for decisions, which can enable us to check for things like fairness and discrimination (Barocas and Selbst 2018). Challenging is a basic feature of our interactions with big institutions, be it governments, universities or banks. Even something as innocuous as a parking fine requires a basic explanation so that we can understand whether the rationale was fair, who was responsible, and whether we have grounds to contest.

[5] The analysis in this section of what it means to explain something, and the different types instrumental and non-instrumental value of explanations, draws from the approach taken in Lazar (2021) and Lazar (2022).

A second reason why explanations are instrumentally valuable is that they allow for improvements to systems when they go wrong (Lombrozo 2011). If we know that a machine made an incorrect decision, for instance diagnosing a mole as cancerous when it was not, but we don't know why it made that decision, we can't intervene to improve it. This is particularly relevant for data scientists who develop and refine AI systems, but also for the business owners of these systems and regulators. Explanations of this type also allow individuals to understand decisions to improve their chances for next time; if a bank denies a loan because of factors including too much debt, individuals can aim to alter their behaviour before reapplying. So, instrumentally, explanations enable decisions to be challenged, they allow systems to be improved, and relatedly, they allow individuals to interact more effectively with those systems.

There are at least two reasons why explanations are non-instrumentally valuable. First, they are an important part of answerability to others as moral equals; denying someone an explanation for a decision you've made that affects them may constitute a denial of their equal moral standing as a human and mutual membership of a moral community (Lazar 2021). If this is right, even a seemingly banal example of a worker denying a colleague an explanation for why they borrowed their mug or were late to a meeting could be a rejection of their answerability to another as a moral equal. This type of explanation makes most sense in interpersonal settings as it seems to require certain mental states from the actor, namely recognising another as an agent of equal moral standing, and acting from a motivation of equality and respect for them as such. It's not clear that organisations could fulfil this. Organisations can give explanations to decision recipients to fulfil other functions, like understanding and challenging decisions, but recognising another as an agent of equal moral standing, and acting from a motivation of equality and respect for them as such, does not seem possible for groups lacking mental states.[6]

[6] This is not to deny that there may be individual people *within* institutions capable of giving explanations of this type. But many of the interactions we have with institutions, whether it be governments or banks, are dealings with them as an entity rather than individuals within this entity. This issue is explored further in the final section of the essay.

A second reason why explanations are non-instrumentally valuable is that they demonstrate that actions were based on the right kinds of reasons. We often want not just a decision, but reasons behind a decision, where the process of deliberation is itself important (Christopher 1998, from Lazar 2021). Acting on the wrong kinds of reasons can itself be wrong; for instance, if our deliberations were based on sexist reasoning, or on information that should have been private – explanations allow this information to come to light. These sorts of reasons could make a decision unjustified, or less praiseworthy, even if it happened to be the right decision. Consider J.S. Mill's case of saving a drowning man in the hope of being paid a reward (Mill 1863). Individuals are also accountable for their intentions and beliefs; for example, the differences between first and second-degree murder, the first being premeditated and the second unintentional and thus not as bad as the first, or the role of mental states in establishing recklessness and negligence (Lazar 2021). So, answerability to others, and demonstrating that one acted for the right kinds of reasons, are two ways in which explanations can realise non-instrumental value.

3. **Why are AI systems opaque?**

Before analysing whether or not AI systems can realise the types of explanation outlined above, it's necessary to first understand why AI systems are said to be opaque, and consider methods to reduce this opacity.

There are at least four forms of opacity in AI systems; institutional, commercial, educational and technical. Institutionally, algorithms are often designed in organisations with input from many engineers and developers over time, meaning "a holistic understanding of the development process and its embedded values, biases and interdependencies" is not possible (Mittelstadt et al. 2016: 7). Commercially, the data and algorithms used in algorithmic systems are often kept secret, backed by trade secrecy protections (Burrell 2016). Educationally, relevant parties might not have the required expertise to understand decision outputs, even if (mathematical or technical) explanations can be given (ibid). However, the concern about AI systems' inability to provide explanations usually relates to the *technical* inability for its outputs to be explained in a way that even an expert would be able to understand. This is what Vrendenburgh (2022) calls "in principle explainability", an issue that is said to afflict only modern machine learning (ML) algorithms.

What makes algorithms technically opaque? Traditional algorithms did not have an explainability problem like those faced by current machine learning systems. This is because in traditional algorithms, rules and weights were pre-specified by the human engineer (Mittelstadt et al. 2016); systems could not do anything that was not already factored into the developers' design for how it should operate given certain inputs.[7] Machine learning, on the other hand, uses vast amounts of data combined with algorithms that can create their own rules to improve their performance. When trained on a certain decision task, like whether to approve a loan, these systems "essentially derive [their] own method of decision making" where it is "simply not known in advance what rules will be used to handle unforeseen information" (Zerilli et. al. 2019, 6). As a result, ex ante predictions and ex post assessments of the system's operations are not possible (ibid). But this ability to find patterns in huge amounts of data is also part of their promise; along with being used in decision contexts where humans used to be, for example diagnosing a cancerous mole, they are also being used in new contexts where human decision makers do not have robust, predictively powerful causal generalisations (Barocas and Selbst 2018); like predicting cancer many years in advance.

A fundamental issue with these models is their detection of correlations rather than causation. Machine learning algorithms often find surprising correlations in huge datasets, but the complexity of these algorithms means it is difficult to pick out a

[7] Though as Zerilli et al (2019) note: "traditional algorithms, like expert systems, could be inscrutable after the fact: even simple rules can generate complex and inscrutable emergent properties. But these effects were not baked in."

smaller set of explanatorily relevant variables and simple relationships between those variables which could explain their decisions (Vrendenburgh 2022). Of course, some of the detected correlations might have causal underpinnings, but the problem is that it's not possible to tell, at least not without detailed further investigation. As Vrendenburgh notes, complexity and an inability to pinpoint causation is not always a barrier to understanding; the natural world is often incredibly complex, but scientists create simplified models to understand it. Yet techniques used to create simplified models in the natural world, like idealisation and abstraction, are not available for these machine learning models because of their complexity.[8] So, while traditional algorithms did not have an in-principle explainability problem, machine learning algorithms do, because of their complexity and their detection of correlation rather than causation. However, there are techniques to limit this in-principle opacity, and so the extent to which opacity matters depends on the kinds of explanations needed in different contexts.

The field of explainable AI has grown significantly in recent years. Techniques have been developed to improve the explainability of decision-making systems, including by making simpler approximations of a model, or by creating counterfactual explanations that show how a model's prediction can be changed by changing one input. The demand for explainability has also led to certain techniques being prioritised in the model development phase, for instance ensuring that training data are correctly labelled and categorised, meaning that the idea of "black box" AI is less well- founded now than it was in the past. McDermid et al. (2021) outline three types of explainability methods. First, for relatively simple models using linear regressions or decision trees (where there is a direct linear relationship between features of the dataset and outputs), understanding how the model works is straightforward; the weights in the linear regression model can be isolated and extrapolated to give insight into the larger model.

[8] Barocas and Selbst (2018) give four reasons why machine-learning models are complex: linearity, monotonicity, continuity, and dimensionality. Basically, this means that ML models don't act in a linear and comprehensible way, allowing them to identify unintuitive and unexpected correlations, but also causing problems for understanding how they have come to their conclusions.

Second, for complex ML models, there are feature importance methods, which involve changing an input feature and observing the difference with the original output. For example, if a model's prediction (e.g. for credit risk) does not change much by tweaking the value of a variable (e.g. age), that variable for that particular data point may not be an important predictor. Third, there are example-based methods, where particular input instances are used to explain complex ML models, for example using counterfactual explanations, adversarial examples or influential instances. In the credit risk example, a counter-factual explanation may ask, "if I had more savings/less debt, would I have been approved?" It's not possible to provide exact causal reasons, or the exact weightings of different factors, but general features of a model providing insight on how it made a decision can be given.

To be sure, there may still be many problems with these techniques to increase explainability. For one, these may be quite technical solutions, appropriate for data scientists, but which might need to be translated into terms that other stakeholders, like regulators, business owners, end users or the public can understand. Explainability techniques can also be expensive and time-consuming; businesses developing AI models may not be incentivised to invest in these methods without government intervention. There is also the problem that more intrinsically explainable models (the "simple ML models" that McDermid et al. outline) are generally less powerful and accurate, so there may be a trade-off between explainability and accuracy. But even for the techniques to improve explainability for more complex models (feature and example-based methods), all that is being provided is a correlative explanation, not a causal explanation. The question then becomes whether causal explanations are indeed needed for work contexts where AI is being deployed.

In what follows, I argue that merely correlative explanations provided by explainable AI are sufficient to achieve the instrumental value of explanations in many cases. However, there are work contexts where the non-instrumental value of explanations are important, and therefore AI should not be used given its inability to provide causal reasons, or reasons that instantiate respect for people as such.

4. AI and the instrumental value of explanation

I now move to the core of my argument, where I make two main claims. First, the instrumental value of explanations - challenging and improving decisions - are generally what's required for decision subjects interacting with large institutions, and these can be provided by current techniques in explainable AI. Second, sometimes explanations tracking non-instrumental value - instantiating respect and acting for the right reasons - will be needed, and these cannot be given by AI. So far, I've mentioned various stakeholders to whom explanations may be owed – decision subjects, data scientists, business owners, regulators, the general public etc. From here on, my focus is primarily on decision subjects; while explanations are important to various stakeholders, decision subjects have the most at stake. They are the ones most directly affected by legal, medical, financial and governmental decisions made about them, and so they have the strongest claim to explanations.

My first claim is that the main reason explanations are important to decision subjects is that they allow decisions to be understood and challenged. Most cases of automated decision making, both now and in the near future, involve large institutions making decisions that affect individuals. While we might be worried about AI being embedded within Spotify or Google to recommend music or websites, the demand for explanation is greatest for decisions that significantly affect our life chances. When we're rejected from a job application, denied a loan, recommended a certain health treatment, refused bail etc. we want to know how and why these decisions were made so that we can contest them if need be and navigate these institutions to achieve our aims in the future. There may be non-instrumental value in understanding why the decision was made, which we may care about even if it doesn’t translate into action, but primarily explanations serve an instrumental good of navigating institutions; what Vredenburgh (2021) calls “informed self-advocacy”. This includes both “forward looking exercises of agency” including understanding rules and procedures of an organisation to achieve one’s goals, whether it be getting a loan or accessing welfare or a certain health treatment, as well as “backwards-looking exercises of accountability” to remedy mistakes or unfairness. Explanations have always served this important function of illuminating our social world and interaction with large institutions (bureaucracies are notoriously opaque) but AI ups the stakes and the challenges.

However, this type of explanation that allows for a general understanding of decisions in order to contest, can be provided by AI systems. As outlined above, methods in explainable AI essentially equate explanations with post-hoc interpretability. That is, they allow decisions to be understood after the fact by identifying general relevant factors that influenced the outcome. Post-hoc interpretability allows for enough information to be gathered to be able to contest decisions and engage in informed self-advocacy. Consider the case of a bank using AI to assess loan applications. Here, general information about how loan decisions are made, for instance credit history, amount of debt, the size of the loan relative to amount of savings etc, would be enough for individuals to assess whether they have been treated fairly or if they have grounds to contest. If, for instance, it was shown that age or postcode was factored into the decision making, then this would be grounds to challenge as it may be discriminatory. Individuals might think they have an adequate credit history, or aim to increase their savings and decrease debts before applying again for a loan. What’s important here is that all that’s required to achieve explanations allowing for contesting or improving decisions are general factors and counterfactual explanations. Causal underpinnings of the decision are not needed for many of our interactions with large institutions.

Now, one might object that institutions owe us specific causal reasons for decisions that seriously impact us. Perhaps causal reasons are necessary to establish why something is unfair, for example if a decision has factored in demographic details like gender, race or age, or why we think the wrong decision was made, like if we think our credit history or savings are adequate. While this seems reasonable in one sense, it would be placing too high a burden on institutions. Even before AI, the right to explanation does not exist in the abstract; it imposes costs on institutions who need to put structures in place so that such explanations can be provided (for example, ensuring information on websites is up to date, that there are adequately trained customer service representatives etc.). Consider acceptance into a university: it is enough to provide applicants with general factors that influenced the decision such as past grades and departmental fit, without detailing specific reasons for individual candidates. The same applies for many other decisions, whether it be applying for a loan or accessing government welfare. And indeed, the counter-factual approach of identifying general factors will allow for things like discrimination to be uncovered; if an institution says that they have factored in age, race or gender, the decision subject can make a judgement of whether this is appropriate. For university admissions, this would be acceptable due to affirmative action considerations, but for loan approvals, likely not due to discrimination law. Maybe in a perfect world we would get specific causal reasons for every decision made about us, but this seems generally unnecessary for the main value explanations provide to decision subjects in navigating institutional rules and procedures.

Another objection asks whether the methods in explainable AI can really provide the types of general-purpose explanations that are needed to understand organisational systems and rules. Instead of identifying general factors that influenced a decision through a counter-factual approach, what is needed is a more holistic overview of the purpose of a decision, how it relates to other decisions and policies within an organisation, and perhaps also mechanisms to contest (e.g. right of appeal services).[9] For people who've had a bank loan application denied, what's needed is not just factors that influenced the decision (savings, credit history etc) but also perhaps some general points about the rationale behind risk calculation for a bank, other services available, where to contest and so on. Current methods in explainable AI are designed to illuminate the technical aspects of a decision tool, rather than to provide such general purpose explanations in natural language.

This seems right, but it just shows that the account so far is incomplete. Explainable AI methods will just be one factor in a bundle of initiatives needed to help individuals understand and navigate institutions. Other factors include general purpose explanations about the purpose of a decision or tool, its contexts, and information about discrimination and fairness, due process etc.[10] Moreover, with advances in natural language processing and text generation, it is entirely possible that these more general types of explanation and information surrounding decisions could be provided by AI. In any case, these are all familiar functions that institutions already have, and the use of automated decision making does not render these factors obsolete, in fact it likely just highlights their importance. So, explainable AI techniques may need to be used in conjunction with more general-purpose explanations so that institutional decision making is understandable and contestable.

[9] Two papers that make a similar point are Zerili et al (2019) who draw on Dennett's theory of 'intentional stance' to say that explanations need to be targeted at the level of practical reason, rather than uncovering the architectural innards of a tool. And Vredenburgh (2021), who calls these 'normative explanations' where the purpose is to communicate the normative reasons for why a decision was made, including possibly reference to organisational policies, without reference to casual reasons.

[10] While this does seem to rely on a decision subject having generally high levels of education and awareness, for example about what counts as discriminatory in loan or hiring decisions, this is a general problem for institutions and not specific to AI. It could be countered by ensuring accessible links to relevant legislation, or simple explainers for what people's rights are. As Vredenburgh (2021) highlights, most rights (whether it to be explanation, privacy or whatever) impose costs to certain parties to ensure those rights are upheld. Organisations have always had structures in place to ensure decisions can be understood and challenged by individuals, and these functions arguably just need to be adjusted and strengthened in an age of AI.

5. AI and the non-instrumental value of explanation

I now argue that there are some contexts where the non-instrumental value of explanations appears important, and this won't be able to be achieved by explainable AI. My analysis so far has focussed on individuals interacting with large institutions: using AI to determine eligibility to a good or service, like welfare payments, a job, entry into university, or recommending a course of treatment like in healthcare. This is the site of much current automation. But there are interpersonal examples, both now and in the future, that fall outside these cases, where the value of explanation is non- instrumental because it evidences decisions were based on the right reasons, or it instantiates respect for people as such. In these cases, the use of AI appears inappropriate if we want to retain the important non-instrumental value that explanations provide.

Consider explanations evidencing that decisions were made for the right kinds of reasons. We often want not just decisions, but reasons behind a decision; acting on the wrong reasons can itself be wrong. For instance, if our deliberations were based on sexist or racist reasoning, or on information that is private, explanations allow this to come to light. We've seen that some of the underlying reasons for AI decisions can be uncovered, for example through counterfactual approaches to explanation that outline the general parameters for how decisions were made. But the fact that this relies on correlation means it still falls short of being able to give causal reasons for decisions. We may know general factors that influenced an outcome (like a loan application), but we don't know specific reasons. This may be acceptable for interactions with large organisations, where the main value of explanation lies in us understanding in general terms why decisions were made so that we can navigate systems and challenge decisions, but it is more troubling in cases where these underlying reasons matter in their own right.

Consider criminal sentencing – here, it's not just a matter of coming to the right decision, but about judges weighing up and giving specific reasons for decisions, balancing factors including the defendant's intention, knowledge and negligence with impacts on other parties. Even if we could develop an incredibly sophisticated legal program that synthesised all aspects of relevant case law, and was able to come up with judgements that were verified by, say, blind reviews by the most experienced judges in a jurisdiction, not knowing the specific causal reasons for judgements would be a problem. One concern is how a defendant could appeal a decision if they don't know the specific legal argumentation and reasons given. But beyond this, there seems to be non- instrumental value in drawing on the right kinds of reasons. We want judges to be able to prioritise certain kinds of reasons, like intention and negligence, while at the same time disregarding other types of reasons, like demographic details or their intuitions about a person. There is non- instrumental value in knowing and being able to evaluate the reasons that a judge gives, independently from the instrumental values this also achieves in being able to challenge decisions. Similarly, in health care, it may be hard to justify a hospital using AI for end of life decisions, even if it could be guaranteed that such decisions were optimal (e.g. best use of resources), if they can't give specific reasons for why a patient's life ended. There is non-instrumental value in knowing the specific reasons why one sick patient was given an organ donation over another, for instance. In such cases involving interpersonal accountability, corelative explanations are not enough.

This leads to a second reason why explanations are non-instrumentally valuable: that they evidence respect for people as such and are an important part of answerability to others as moral equals. Denying an explanation for a decision you've made that affects someone else could constitute a denial of their inherent worth as a human and mutual membership of a moral community. Here, following Lazar (2021), the idea is that part of answerability to others as moral equals is being responsive to them, and giving them explanations when decisions are made that significantly affect them. These types of explanation are not just about justifying one's actions, but about explaining oneself from a motivation of equal regard for another.

But is this right? Even if we accept that this kind of answerability is important for many interpersonal relationships, it's less clearly needed in many work contexts dictated by rules, procedures and risk management protocols. In many jobs, it's enough that workers perform their basic functions well. The institution as a whole should be able to justify its practices with the right kinds of explanations. But it would be unfair to expect this from each worker. A nurse is required to explain what medicine they're giving to a patient because it's hospital procedure to explain to a patient what treatment they're being given. A bank teller is required to explain to a customer what they've done with their money because it is company policy. There may be underlying reasons for such policies, such as consent or legal liability, which may ultimately be grounded in something foundational like respect for persons, but the reason why workers provide such explanations is because of procedure. Much of work life is dictated by rules and policies, and so even though it would be nice to retain this non- instrumental value of explanation, it's not strictly necessary.

However, there are examples, like the case of determining criminal liability, where these sorts of explanations are constitutive of the good in question. Part of natural justice arguably involves being seen and tried by another member of one's moral community. There is something of value in being tried by a human judge, even if they are biased or coldly rational, because they are part of one's moral community and share similar capacities to us. Similarly, in certain types of life-or-death medical decision making, such as determining which sick patient to give a blood transfusion or kidney transplant, it matters that patients or their families are given explanations from another being within their moral community who can be seen to relate to their plight. Given the moral weight of these decisions, it seems important not just that an explanation is provided about the rationale for such decisions, but that this explanation comes from a person capable of relating to them as an equal. Of course, there might be cases where we choose to sacrifice this value, say if a fully automated hospital would save considerably more lives than a non-automated hospital. But such a situation would incur moral costs.

Again, it might be objected why being seen by another member of one's moral community is really that important. There are various ways of justifying this idea. One route is to say that certain institutions are constituted by human relationships and practices which do not seem amenable to replacement with machines. Pasquale (2018), for instance, argues that types of automation within the legal system threatens the "deliberative governance" that is a foundation of a "just and accountable social order." Another approach is taken by Danaher (2019) who argues that the aggregate effect of widespread automation will atrophy and degrade the bonds of human interdependence that give our lives shape and meaning. It is outside the scope of this essay to argue why being seen by another human is important in certain contexts, but if we accept even a minimal form of this intuition, it appears to place constraints on the use of AI if we are to account for the non-instrumental value of explanations.

Overall, I have argued that in most cases of automated decision making, the main reason why explanations are important to decision subjects is that they realise the instrumental goods of enabling people to challenge decisions and navigate institutions to achieve their aims. However, there are limits to this: in some cases, it appears that explanations have non-instrumental value in that they evidence that decisions were made for the right causal reasons, or that they instantiate respect for people as such. This analysis approach can be used to make sense of individual cases where there is a demand for AI explainability. First, we can ask: what is the value of explanations in this context? I have outlined two instrumental, and two non-instrumental values in this essay, but there may well be more. Once these values have been identified, we can ask: can AI achieve these values? Taking medicine as an example, explainability methods can help identify general factors about a medical diagnosis or treatment recommendation, which allows us to navigate a hospital system and challenge if necessary. But in cases of end-of-life care, explanations that give specific reasons for decisions or that evidence respect for people as such seems necessary, and therefore AI would be inappropriate in such contexts. This basic framework can be applied to many other cases in the same way, whether it be explanations in the legal system, governments or corporations.

A final caveat: as alluded to in section three, there are commercial and institutional reasons why explainability methods may not be used, even if they are available. Namely, these methods are expensive and time-consuming, so companies or organisations wanting to deploy their AI products as soon as possible may forgo implementing them without some external incentive like regulation. The purpose of this paper is to show how, in theory, the types of explanations often needed can be provided by currently existing technology. Whether these will actually be used in practice is a matter of incentives, politics and regulation.

Conclusion

Explainability techniques can give a general indication of the parameters for why an AI decision was made. But it can't give specific causal reasons. I have argued that in many cases, the value of explanation lies in the instrumental value of being able to challenge and contest decisions that are made about us. But in some cases, especially in high stakes decision making such as determining criminal liability or end of life care, specific (causal) reasons really matter. In such cases, the use of AI appears inappropriate given its detection of correlation rather than causation, and the fact that machines are not fellow members of our moral community. By clarifying the instrumental and non-instrumental value of explanations, and applying this to AI as it is being deployed in work contexts, I have argued that the demand for AI explainability can often, though not always be met.

References

Barocas, S., Selbst, A. D. (2018). ‘The Intuitive Appeal of Explainable Machines’, *Fordam Law Review*, 87 (3): 1085–139.

Booth, R. (2019). ‘Benefits system automation could plunge claimants deeper into poverty’, *The Guardian*, 14 October. Available at: https://www.theguardian.com/technology/2019/oct/14/fears-rise-in-benefits-system-automation-could-plunge-claimants-deeper-into-poverty (Accessed 1 September 2022).

Burrell, J. (2016). ‘How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms’, *Big Data and Society*, 3: 1–12.

Christopher, R. (1998), 'Self-Defense and Defense of Others', *Philosophy & Public Affairs*, 27 (2), 123- 141.

Crawford, K., Schultz, J., (2014). 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms', *Boston College Law Review*, 55, 93-128.

Danaher, J. (2019). ‘The rise of the robots and the crisis of moral patiency’, *AI and Society*, 34 (1):129- 136.

Doshi-Velez, Finale, et al. (2017). 'Accountability of AI under the Law: The Role of Explanation', arXiv:1711.01134.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). ‘Human decisions and machine predictions.’ *The Quarterly Journal of Economics*, 133(1):237–293.

Lazar, S. (2021). ‘The Value of Explanations’, manuscript available from author on request.

Lazar, S. (2022). ‘Legitimacy, Authority, and the Political Value of Explanations’, keynote address to Oxford Studies in Political Philosophy workshop. Available at: https://philpapers.org/archive/LAZLAA-2.pdf.Lombrozo 2011

Lee, K. (2019). ‘The Third Revolution in Warfare’, *The Atlantic*, September 11. Available at: https://www.theatlantic.com/technology/archive/2021/09/i-weapons-are-third-revolution-warfare/620013/ (Accessed 1 December 2021).

Malle, B. F. (2004). How the Mind Explains Behaviour: Folks Explanations, Meaning, and Social Interaction. Cambridge, MA: *MIT Press*.

Medhora, S. (2019). ‘Over 2000 people died after receiving Centrelink robo-debt notice, figures reveal’, *ABC News*, 18 February. Available at: https://www.abc.net.au/triplej/programs/hack/2030-people-have-died-after-receiving-centrelink-robodebt-notice/10821272 (Accessed 20 June 2022)

McDermid J., Yan, J., Porter Z., Ibrahim, H., (2021). ‘Artificial intelligence explainability: the technical and ethical dimensions’, *Philosophical Transactions of the Royal Society*, 379(2207).

Mill, J. S., (1863). Utilitarianism, London, Parker, son, and Bourn.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., Floridi, L. (2016). ‘The ethics of algorithms: mapping the debate.’ *Big Data and Society*, 16,1–21

Müller, C., (2021). ‘Ethics of Artificial Intelligence and Robotics’, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/>.

Pasquale, F. (2018). ‘A Rule of Persons, Not Machines: The Limits of Legal Automation’, *Maryland Faculty Scholarship*. 1612.

Vredenburgh, K. (2021). ‘The Right to Explanation’ *The Journal of Political Philosophy*, 30(2):209-229.

Vredenburgh, K. (2022). ‘Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace’, *Canadian Journal of Philosophy*, 52(1):78-92.

Wilkenfeld, D. (2014). 'Functional Explaining: A New Approach to the Philosophy of Explanation,' *Synthese* 191:14.

Yala, A., Strand, F., Smith, K., (2021). ‘Toward robust mammography-based models for breast cancer risk’. *Science Translation Medicine*. 13(578).

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). ‘Transparency in algorithmic and human decision-making: Is there a double standard?’, *Philosophy and Technology*, 32(4):661–683.