Ethics and Information Technology (2022) 24:49
https://doi.org/10.1007/s10676-022-09654-x

ORIGINAL PAPER

Explanation and Agency: exploring the normative-epistemic landscape of the "Right to Explanation"

Fleur Jongepier[1] · **Esther Keymolen**[2]

Published online: 11 November 2022

Abstract
A large part of the explainable AI literature focuses on what explanations are in general, what algorithmic explainability is more specifically, and how to code these principles of explainability into AI systems. Much less attention has been devoted to the question of why algorithmic decisions and systems should be explainable and whether there ought to be a right to explanation and why. We therefore explore the normative landscape of the need for AI to be explainable and individuals having a right to such explanation. This exploration is particularly relevant to the medical domain where the (im)possibility of explainable AI is high on both the research and practitioners' agenda. The dominant intuition overall is that explainability has and should play a key role in the health context. Notwithstanding the strong normative intuition for having a right to explanation, intuitions can be wrong. So, we need more than an appeal to intuitions when it comes to explaining the normative significance of having a right to explanation when being subject to AI-based decision-making. The aim of the paper is therefore to provide an account of what might underlie the normative intuition. We defend the 'symmetry thesis' according to which there is no special normative reason to have a right to explanation when 'machines' in the broad sense, make decisions, recommend treatment, discover tumors, and so on. Instead, we argue that we have a right to explanation in cases that involve automated processing that significantly affect our core deliberative agency and which we do not understand, because we have a general moral right to explanation when choices are made which significantly affect us but which we do not understand.

Keywords deliberative agency · AI · epistemic goods · automated-decision making

Introduction

If machines or algorithms make a decision that has significant consequences for your future or daily life, like a medical AI system diagnosing you with cancer or deciding on a personalized treatment, then it's natural to think that the least you're entitled to is an explanation of why that particular decision was made and how. This normative intuition seems to be deeply ingrained both in much of our public life and medical contexts in particular, as well as in recent academic discussions on explainable AI (e.g. Pasquale, 2015; Miller, 2017; Selbst and Barocas, 2018; Mittelstadt, Russell, and Wachter, 2019) and in legal debates about the so-called "right to explanation" in the European General Data Protection Regulation (e.g. Goodman and Flaxman, 2016; Selbst and Powles, 2017).

A large part of the explainable AI literature focuses on what explanations are in general, what algorithmic explainability is more specifically, and how to code these principles of explainability into AI systems (Gilpin et al., 2018, Beaudouin et al., 2020). The existing legal literature on the right to explanation has predominantly focused on the factual question whether there *is* a right to explanation (implicit) in the GDPR or not. Much less attention has been devoted to the question of *why* algorithmic decisions and systems should be explainable and whether there *ought* to be a right to explanation and why. An answer to that more fundamental moral question would be helpful in the context of

✉ Fleur Jongepier
fleur.jongepier@ru.nl

Esther Keymolen
e.l.o.keymolen@tilburguniversity.edu

[1] Radboud University Nijmegen,
9010, 6500 GL Postbus, Nijmegen, The Netherlands

[2] Tilburg Law School, TILT, Tilburg University,
90153, 5000 LE Tilburg, The Netherlands

discussions about explainable AI, the GDPR, and medical AI.

In this article, we therefore explore the normative landscape of the need for AI to be explainable and individuals having a right to such explanation. This exploration is particularly relevant to the medical domain where decisions generally have a fundamental impact on a patient's chances of living a good life. Explanations in these contexts thus often matter substantially. The (im)possibility of explainable AI is high on both the research and practitioners' agenda and the recent proposed AI Act, which identifies AI systems in medical devices as high risk, will probably make sure that it remains there for the foreseeable future. Rapid developments in AI and Machine Learning lead to new diagnostic tools, enable personalized treatment, and improve or take over tasks that were previously executed by clinicians (e.g. medical imaging analysis). The often inherent opaqueness of these AI-powered systems has been seen as a possible threat to trust in the given diagnosis, in the doctor-patient relation, and in the medical system at large (Hatherley, 2020). While there is some debate on the specific role of explainability, for instance if it should always trump other values such as predictive and diagnostic accuracy (London, 2019), the dominant intuition overall is that explainability has and should play a key role in the health context.

Notwithstanding the strong normative intuition for having a right to explanation, intuitions can be wrong. So, we need more than an appeal to intuitions when it comes to explaining the normative significance of having a right to explanation when being subject to AI-based decision-making. The aim of the paper is therefore to provide an account of what might underlie this normative intuition.

The paper consists of two parts. In the first part, we address the question of whether there are *special* normative reasons for having a right to explanation in cases involving automated decision-making that involve no human intervention, both inside and outside of medical contexts. Those who answer this question positively accept what we refer to as the 'asymmetry thesis', according to which automated and human decisions each introduce different normative challenges and make different normative claims on us. We argue, however, that the asymmetry thesis rests on a mistaken conception of the relation between technology and human agency. Namely, one according to which human agency is defined as being independent of or opposed to technology, and vice versa.

We argue that the asymmetry thesis not only underestimates the prima facie non-problematic nature of human decisions but also overestimates the prima facie problematic nature of non-human decisions. In many cases, the fact that 'machines'make decisions that affect us without our understanding how the decision was reached, is no immediate cause for concern. [1] After all, technology often enhances and extends our agency rather than posing an obstacle or standing in opposition to it. We thus argue that the asymmetry thesis should be given up, if only because we don't need to appeal to such a view in order to explain the normative intuition of why we have a right to explanation.[2] In its place, we defend what we call the 'symmetry thesis', according to which there is nothing normatively special let alone prima facie bad or wrong about automated decisions as such. This is compatible with automated decision-making often being bad, wrong, opaque or unintelligible *in practice* – indeed that is what we will be arguing for, too. The right to explanation would then help guarantee that this intertwinement is of the right sort and that the network of human and non-human actors involved in automated decision-making is balanced in the right way.

In the second part of the paper, we turn to the question of what normative reasons might then underlie having a right to explanation, if not an appeal to the simple fact that one is subjected to a decision in which (supposedly) no humans were involved. We propose to answer this question by exploring the connection between having the ability to understand important decisions made about a person and what we call a person's 'deliberative agency': in a nutshell, a person's ability to formulate and act on her own reasons.[3] Irrespective of whether humans or machines were the main driving force behind a decision, a person has a general right to explanation when her agency is substantially affected or the vital means for agency are undermined. Something along these lines has been mentioned in the literature (Selbst and Powles, 2017), indeed connections between 'explanation' and 'autonomy' are often vaguely made, but have not been sufficiently worked out so far. This connection between deliberative agency and a right to explanation is particularly interesting —and challenging— in the doctor-patient relationship, which is characterized by an inherent information asymmetry (Goodyear-Smith and Buetow, 2001). For instance, depending on the model underpinning this relation —ranging from paternalistic ones where the doctor makes all decisions on their own to informed-decision making ones where the patient receives all information and gets to

[1] For readability, we'll refer to 'machines' to include algorithmic, data-driven systems, including those that use AI and ML techniques.

[2] In what follows, we shall use the phrase "have a right" to specify a moral claim, i.e. the phrase is not used descriptively as referring to legal articles.

[3] This notion is closely connected, and indeed on certain interpretations identical, to the concept of 'autonomy'. Given that 'autonomy' has been defined in highly diverse ways, including conceptions that are quite different from what we have in mind with the concept of deliberative agency (such as higher-order identification with certain desires or non-oppression), we choose to operate with this more specific term.

decide—a right to explanation can take on relevantly different shapes. This is also evident from discussions about informed consent in medical ethics, hence we propose it may be fruitful to approach the right to explanation from a parallel angle.

The right to explanation gets its normative force, we argue, from the fact that knowledge and understanding are necessary conditions for deliberative agency. The normative intuition that we have a right to explanation is thus not about automated decisions as such, but more generally about how decisions affect or undermine our deliberative agency, combined with the observation that automated decisions do so more often, or that they are more likely to pose a threat to deliberative agency.

Humans and Machines: the Asymmetry Thesis

Why would we have a right to meaningful information when we are made subject to automated processing decisions, or when a physician chooses to rely heavily—all too heavily—on AI systems? What justifies this normative intuition? A right to meaningful information could be grounded in the intuition that when (medical) machines make decisions that affect us special normative considerations are generated. In other words, when an AI system provides us with a diagnosis or recommends a certain treatment, we are entitled to receive an explanation why and how such a conclusion has been reached. We have a right to this explanation, the thought goes, because machines and not human beings make substantial decisions about us. The underlying assumption is thus something like the following: when it is an automated, data-driven system that comes to a decision related to my health and well-being, I have a right to know. The grounds of a right to explanation first and foremost concern the who or what is doing the decision-making, not the content of the decision.

This way of answering the question about why we have a right to explanation amounts to what we will call a normative "asymmetry thesis". This thesis holds that from a normative standpoint, automated decisions are valued as normatively different or special in comparison to human decisions. In short, in order to have a right to explanation, it makes a difference whether a doctor diagnoses cancer in a patient, versus an algorithm coming to the same conclusion. Notwithstanding the same outcome, the process leading up to it is perceived as being fundamentally different, so different that it leads to different normative conclusions, namely, whether or not we have a right to explanation.

It's unclear how widespread the normative asymmetry thesis is, and who – either in academia or society at large – would subscribe to it, implicitly or explicitly. We do not aim to make a descriptive claim about the pervasiveness of the asymmetry thesis. Rather, our point is that it is the debate concerning the right to explanation, explainable (medical) AI, as well as calls for opening up the black box nearly all start from the assumption that we have such a right: that AI ought to be explainable, and all black boxes must be opened. In other words, the question of *why* gets much less attention (though see Ananny and Crawford, 2018), perhaps because theorists think the answer is obvious, whereas we don't believe it is.

Importantly, the asymmetry thesis is a natural normative counterpart of particular metaphysical understanding of 'human' versus 'automated' decisions. On the strongest version of this conception, the one is to be understood in terms of the *absence* of the other. In other words, a human decision on this view is a decision without the intervention of automated processes; an automated decision is a decision without the involvement of a human actor.

However, it's naïve to think that there actually exists a pure, automated decision in the sense of a decision where no human beings were in the loop. Bruno Latour (Latour, 1992, 1993) famously explained that when technologies become part of our everyday practice, we tend to forget that they actually consist of a network of different human and non-human actors. By focussing only on the output of a technology (the decision), we no longer take into account that this outcome is actually the interplay of a variety of associations of engineers, algorithms, data scientists, insurers, medical experts, hardware, corporations, software, regulators and other stakeholders. Moreover, Latour (1993) speaks of a 'generalized symmetry' to stress the interconnectedness of humans and non-humans. The underlying assumption is that to understand technology or human beings, we have to focus on the network that connects them. Many of the actions of human beings can only be properly understood when taking into account the technology that enables (or disables) them to act in the way that they do.

Although the asymmetry thesis and the accompanying metaphysical framework is perhaps rarely explicitly embraced in the domain of philosophy of technology, it may still be implicitly held in academic debates, including the medical debate, and it's *not* unlikely that the view is even popular outside of these domains. We believe there is value in clarifying why it must be rejected.

To start, it is often genuinely unclear what underlying conception of a human or automated decision is adopted exactly. This ambiguity becomes conspicuously clear if we for instance turn to the legal debate. Looking at some of the vocabulary that is used in the GDPR, such as Article 22's opening sentence: "The data subject shall have the right not to be subject to *a decision based solely on automated*

processing (own emphasis), including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her", or having a right to "obtain human intervention" (Article 22; Recital 71), the wording of these phrases suggest that an automated decision is one in which human beings (or corporations) have had no significant role in the way in which a decision has been reached and how it is applied to a data subject. In practice, though, humans are 'in the loop,' at least somewhere. Following the GDPR, it seems that even extremely minimal human involvement would make a decision qualify as non-automated hence "human".[4] Similar issues arise for the notion of "meaningful human control" which is frequently used by policy-makers and technical designers in debates on the ethics and regulation of autonomous systems. As Filippoo Santoni de Sio and Jeroen van den Hoven point out, we currently "lack a detailed theory of what "meaningful human control" exactly means" (2018).

Also, recently, Sven Nyholm (2018) has called attention to the fact that in debates about self-driving cars and military robots, a great number of theorists still hold onto a naive conception of the type of agency of such automated systems would have, namely, as being capable of acting on their own, independently of any human beings. Nyholm rightly points out that the agency of such systems - however advanced and sophisticated they may be – is best thought of as "a kind of *collaborative* agency—even if the [system] might be doing "'most of the work."'" and where "the humans involved initiate, supervise, and manage" the task (Nyholm, 2018, p. 1211).

Even with self-learning systems and the development of AI, somewhere in the process there will be human beings deciding on which training data is being fed to the system and which parameters and thresholds are being used. Even when we would end up in a context where the algorithms are "in charge", behind them there will always be human beings who have put them there in the first place. Trying to map how a network of human and non-human actors facilitates or hinders certain actions, or comes to a certain decision, allows us to question not only the outcome of such a decision, but importantly it also enables us to critically reflect upon the distribution of responsibility between all these human and non-human actors (Noorman, 2021).

Rather than speaking of automated decision *versus* human decisions, we propose that it is more accurate – and normatively more sensible - to speak of dominantly automated decisions or of dominantly human decisions. How to understand the addition of 'dominantly' will depend either on which of the parties has most responsibility or who has done most of the work. The key question then becomes whether there's an adequate division of labour, communication and justificatory links from one to the other.

In order to establish if such an adequate division exists, it will be necessary to genuinely scrutinize the human-technology relations and analyse how all components relate to each other. A fruitful starting point for such an analysis of human-technology relations can be found in postphenomenology and theory of mediation (Rosenberger and Verbeek, 2015; Aagaard et al., 2018; Ihde, 1990). In this subdomain of philosophy of technology, distinctions are being made in the way in which technologies mediate the interaction of human beings with the world (e.g. a pair of glasses enhances the ability of a person to see the world, a barometer allows a person to read the world in a certain way). While such an analysis can only be done for specific technologies – a pair of glasses mediates a person's behaviour in quite a different way than a smart phone does– generally we can say that data-driven, AI applications lead to what Verbeek (2008) calls "cyber intentionality". It is not just that we use or delegate certain actions to AI, but AI applications have – to a certain extent – some freedom to shape our intentionality. Human intentionality can therefore not be understood anymore without relating it to its technological counterpart. When this kind of cyber intentionality comes into play, the role of technology becomes more apparent and dominantly automated decision-making more plausible. Or as Wellner and Rothman (2020, p. 199) put it, in these cases "the human intentionality 'withdraws' and the technological intentionality 'takes over'". This kind of in-depth analysis of how human-technology relations take shape will be instrumental to come to distinguish between dominantly automated decisions or dominantly human decisions.[5]

Curiously, accepting the idea of fused and collaborative human-technology relations automatically puts pressure on the normative asymmetry thesis. After all, once we abandon the strict distinction between human- and technological decision-making such as outlined above, the supposed special normative considerations no longer follow automatically. This is because it is not strictly speaking true that "*computer* says no". There are always human beings or organizations that *simultaneously* say no, or who have let the computer say no, and who are to be held responsible for their choice to delegate. If it is more appropriate to speak

[4] As Wachter rightly observes: "The outcome may be that robotic decision making would not qualify as "solely" automated. Ironically, this reluctance could make systems less accountable by preventing the GDPR's safeguards from applying." (Wachter, Mittelstadt, and Floridi, 2017)

[5] How to determine in any given case when to speak of dominantly human or dominantly automated decision making is unfortunately out of the scope of this article, but beyond doubt a topic that needs further investigation. We foresee that discussions on distributed responsibility, liability, and accountability in the domain of AI as well as theory of mediation and postphenomenology could be a fruitful starting point for such an endeavor.

of dominantly automated decisions or of dominantly human decisions, then this raises the question: what difference does it make, from a normative standpoint, whether an algorithm instead of a human being is dominantly in charge?

When we consider everyday scenarios, we delegate decisions to automated systems all the time. Algorithms are dominantly in charge when it comes to the songs we listen to (Spotify) or how we get from a to b (navigation systems). So, the mere fact that algorithms or machines decide for us and not human beings is not what makes these decisions prima facie problematic. In fact, in many cases we actually prefer automated systems rather than human beings to make those decisions for us. The idea that a human being would tell us which song to listen to or how to get to our destination would be quite annoying, to say the least.

Even when it concerns decisions that have a serious impact on our lives, such as decisions involving a diagnosis or treatment it does not become immediately clear why all normative work is being done by machines that are (dominantly) responsible for a certain outcome. Assume that an overworked and non-empathetic doctor gives you a call and without any further explanation tells you that you are diagnosed with heart problems and need to undergo surgery. Why would decisions made by arrogant, ignorant, and impatient human beings be prima facie less problematic than decisions made by machines? Why would we have a right to explanation in the latter case and not in the former case, or at least not with the same sense of urgency? From a normative standpoint these two cases are symmetric. We therefore want to suggest we need to accept the 'symmetry thesis' instead. According to the symmetry thesis, what makes a decision problematic is not who or what makes the decision but simply which decision was made and why, i.e. it has to do with the kind of decision that is being made and the justification of the decision.

Acknowledging the merging and collaboration between human and technological actors actually provides us with an extra reason for having a right to explanation. With a right to explanation, we would as it were have a right to summon the human behind the machine to reveal themselves. After all, to have a right normally presupposes that, somewhere down the line, there is some moral agent on whom you can make a claim, in other words, an agent under obligation. That is why Facebook's founder Mark Zuckerberg - and not his algorithm - was summoned before the U.S. congress and the EU commission.[6] Algorithms are always connected to human actors, often intricately so. Thus, even if a health professional would be completely dependent on an AI application to come to a diagnosis, there are supervisory authorities who bear some or a lot of responsibility, there are data scientists and ML experts who developed the model and overlooked the data collection and processing, there is the company these professionals work for —if it is not an in-house developed application— as well as the managers and board of the hospital who decided on the introduction of the machine in the first place—these actors could in principle all be asked to explain their decisions. A right to explanation may strengthen the position of patients as well as of practitioners—who in the end are also in need of information to properly embed these machines in their professional conduct and are often time the most important point of contact for the patient— by ensuring that this intertwinement of human beings and automated systems takes on a reliable and justified form.

A General Moral Right to Explanation

In the previous section, we concluded that we must reject an asymmetry thesis according to which some metaphysical distinction between human and automated decisions forms the key to the answer to the question regarding the normative intuition. We believe some version of a symmetry thesis is more plausible. This interim conclusion, however, actually poses an immediate challenge for our main goal, which was to explain and provide justification for the normative intuition regarding a right to explanation. If there is no metaphysical difference that we can rely on to provide the answer, then it seems we have made the normative landscape surrounding the right to explanation more, not less, mysterious. We believe that this challenge can be met, however. What we want to consider in this section is the 'metaphysically agnostic' strategy, which involves deriving the specific right to explanation in the context of AI- driven practices from a general moral right to explanation.

For this argumentative strategy to be successful, two prior argumentative steps must be made. First, it must be made plausible that there is a general moral right to explanation. Second, it must be shown that in certain contexts, such as the health context, where machines are becoming dominantly involved in making decisions about us, this right is at risk of being violated.

To begin, if there is such a thing as a general moral right to explanation, then this means that we have a right to certain epistemic goods, given that 'explanation' is an epistemic concept. What constitutes an explanation is a vexed

[6] The fact that Mark Zuckerberg is summoned to testify and not his algorithms is because currently these algorithms are not able to provide a sufficient explanation and probably, they will never be. This is however still a contingent fact. Following the symmetry thesis, if at a certain moment in time, in the human-technology collaboration, a technological agent turns out to be better equipped at providing explanations for certain decisions, we might well want to demand explanations from the technological agents involved.

question in the philosophy of science and epistemology (cf. Miller, 2018). Plausibly, in the current context receiving an explanation in the relevant sense includes at least two components. First of all, having an explanation involves acquiring knowledge and understanding. The distinction between the two, very roughly, is that knowledge "is concerned with *propositions*, whereas understanding usually isn't, at least not directly" (Pritchard, 2009, p. 30). What is distinctive about understanding, as Jonathan Kvanvig puts it, is that understanding "has to do with the way in which an individual combines pieces of information into a unified body" and the way in which an individual is able to grasp the relations *between* items of information (Kvanvig, 2003, p. 197).

Second, having a right to explanation in the relevant sense means having a right to a *useful* or *meaningful* explanation. But what makes an explanation 'useful' to a subject? Here, we want to explore the idea that certain epistemic conditions are necessary for what we call a person's 'deliberative agency'. A deliberative being, as we see it, is a rational being in the sense that she has reasons for believing what she does, and reasons for acting in the way that she does. When confronted with evidence, a deliberative agent is able to weigh reasons pro and con, and come to a judgment or plan of action. To be a deliberative agent means that we of course have all kinds of inclinations and impulses, but we have the capacity to take a critical distance from them. The point is not a descriptive one about how we actually, as a matter of psychological fact, reach conclusions or propel to action. It might well be that we quite often act on impulses and inclinations without much reflection (Arpaly, 2003). To be a deliberative agent in the sense at issue here, then, is compatible with claims to the effect that we are often *not* rational in the more common use of that term (cf. Kahneman, 2013).

More positively, being a deliberative agent involves making plans for the future (e.g. (not) to have kids, to undergo or refuse a certain treatment, or to opt for a second opinion); plans that we can reconsider and abandon, if we think we have good reason to do so. These reasons might turn out to be silly or irrational on closer inspection, but they are our *own* reasons. We therefore want to operate with the following conception of deliberative agency: to be a deliberative agent is to be able to come to judgments on the basis of reasons, and to formulate and act on plans on the basis of one's own reasons.

It is important to note that in the literature on explainable AI oftentimes a quite narrow distinction is being made between explanation and understanding, where the former solely has to do with conveying information and the latter with operationalizing this information and relating it to a wider background understanding (Adadi and Berrada, 2018). This is, beyond doubt, a relevant distinction as something can be explainable without being understood (Wadden, 2021, p. 3). We, however, explicitly choose to adopt a broader conceptualization of the concept by including the components of knowledge, understanding, and meaningfulness.[7] In this article we focus on explanations that also lead to understanding as it is only this sort of explanation (rather than, say, receiving some technical code) that is of the sort that would allow people to contest certain decisions. In other words, it is this richer conception of explanation that is relevant for deliberative agency.

Epistemic conditions

The next step of our argument involves showing that conditions need to be in place for one to be (and continue to be) a deliberative agent, and that amongst these conditions are epistemic conditions. That there are conditions that are required to be a deliberative agent is beyond dispute (though a lively debate is still being pursued as to what those conditions are exactly, see e.g. Buss and Westlund, 2018; Mackenzie and Stoljar, 2000). One condition that is accepted by all agency/autonomy theorists for instance is that certain minimal intellectual and/or linguistic capacities are required for the development and sustainment of deliberative agency. It is also generally agreed that 'negative freedom' is required in order to be a deliberative agent: it is necessary that other people or institutions impose no obstacles for one's planned course of action (i.e. not being imprisoned, assaulted, but many agree it also involves 'inner' conditions involving the absence of oppression, manipulation, brainwashing, propaganda etc. (Oshana, 2014; Coons and Weber, 2014).

Importantly, these facets of deliberative agency do not entail that when engaging in this weighing and formulating of reasons, one has to do this in splendid isolation. On the contrary, deliberative agency is importantly social and is made possible and is strengthened by conversations and discussion with others who, for instance, have relevant knowledge or who, due to their intimate relation with this person, can provide her with refreshing perspectives and advice. An important requirement for this form of interactive deliberative agency is that the interlocutor should have the best interest of the agent at heart. Moreover, as stated above, the interaction should be free from manipulation and coercion. In the doctor-patient relation, where fiduciary duties are bestowed on practitioners, entailing that they should always act in the interest of their patients, a form of deliberative *co*-agency can successfully emerge. The practitioner as a

[7] We are thus sympathetic to the multi-faceted approach of Miller (2018), who argues that the field of explainable AI can benefit from considering conceptualizations from philosophy, cognitive science and social science that have a long history of trying to define the notion of an 'explanation'.

fiduciary and ally enables the patient to exercise her deliberative agency by providing and discussing the relevant information and options at hand, ideally reaching a shared conclusion with which the patient can wholeheartedly identify herself.

Now we want to explore a condition that is typically overlooked in these discussions, namely, epistemic conditions.[8] Our basic idea is that in order to make decisions at all, certain minimal epistemic conditions pertaining to knowledge and understanding need to be in place. Let's begin with a simple, everyday life example. In order to decide to take the train rather than the plane to Amsterdam, requires among other things knowledge of what trains and planes are, understanding of how the two differ, knowing something about how long it takes with each mode of transportation to get to Amsterdam, what each mode of transportation costs, and may also require knowing certain facts about climate change. Only if these epistemic conditions are in place, can one deliberate about the options and only then one can be said to be in a position to make a decision. But if one fails to know or understand information that is crucial to making a decision, one lacks the epistemic goods required for successful deliberative agency.

When it comes to the nonfulfillment epistemic condition of deliberative agency, there are intuitively problematic and unproblematic cases. There are of course lots of things we don't know about the world and ourselves that would greatly improve our deliberative agency, but this does not mean we have a *right* to such epistemic conditions, let alone that any moral violations are at issue. For the nonfulfillment of epistemic conditions to take a morally problematic form, we propose that minimally two conditions must be met.[9]

First, there must be other agents and/or institutions (corporations or states) who can be considered responsible for the agent's lack of access to the relevant epistemic goods (or who *ought to have had* access to those goods). For example, there is a lot we do not know about Higgs particles, but no one can sensibly be said to be responsible let alone blameworthy for this epistemic loss. If, however, patients would undergo surgery but not know about their chances of full recovery, the costs involved with their treatment, and the possible discomfort they might experience, then it makes sense to think a practitioner or hospital would be responsible and potentially blameworthy. The former examples do not belong to the relevant categories of responsibility and blame, whereas the latter do. In order to have a right to explanation, there must be some party whom to turn to for the reason that they are the (only or principal) party to turn to regarding the epistemic goods concerned.

Second, the relevant epistemic goods must be substantial, not trivial. The missing knowledge and understanding in question must, in other words, be important to one's beliefs or goals, and one's beliefs and goals must be important to one's identity or way of life. For example, not knowing that the roundabout around the corner is temporarily blocked, the result of which my deliberative agency is impoverished (after all, the lack of knowledge makes it impossible to adapt my travel plans and not come too late to a meeting) is an epistemic loss of a trivial kind. This is so even if condition 1 was met, that is, some person or institution was in the position to inform me about the blockage or intentionally refrained from doing so. Some knowledge is merely facilitative to one's deliberative agency, not crucial to it. In short, our proposal is that we have a right to epistemic goods for what we might call one's *core* deliberative agency, not one's *entire* deliberative agency, including trivial decision-making.

Now if (non-trivial) epistemic goods are necessary for (core) deliberative agency, then we must ask: what sort of epistemic good might 'getting an explanation' be? Epistemic goods come in different shapes and sizes: there is propositional knowledge, but also embodied knowledge (e.g. proprioceptive information about the whereabouts of one's limbs 'from the inside'). But there are also a variety of methods to acquire these different epistemic goods. One might acquire knowledge by surfing on the internet, by conducting experiments, by undergoing psychological tests, but also via introspection and through testimony by talking to friends, family or experts.

Explanation: knowledge of reasons

Getting explanations is a core epistemic good. The concept of 'explanation' is notoriously difficult to pin down, however, and takes on many different meanings and has a rich history in the philosophy of science. The sense of explanation that is crucial for deliberative agency is receiving and having a right to *knowledge of reasons*.[10] Knowledge of reasons is a type of knowledge often acquired in second-personal social encounters, such as dialogues (cf. Eilan, 2014; Heal, 2014). One typically acquires knowledge of (another's) reasons by asking a "why"-question (Anscombe, 1957).[11]

As Daniel Dennett has pointed out, 'why'-questions are ambiguous (Dennett, 1981). Asking 'why' can either involve a "how come?" question, the answering of which involves

[8] We turn to the diagnostic question (of why this might be so) below.

[9] These are necessary conditions, not also sufficient conditions.

[10] NB it is crucial, but not the only type of explanation that is necessary. For instsance, as a patient one might require an explanation in terms of reasons *as well as* more technical or evidence-based information.

[11] We thus take a rather different view than e.g. Pearl and Mackenzie (2018).

citing things like causes, cognitive processes or our habits and a "what for?" question, which requires the citing of reasons. For instance, when we ask Frederik why he orders a vegetarian dish at a restaurant, he could either say he always orders that dish there or that his parents were vegetarian. He could, if he's a neuroscientist, perhaps even say something about how his brain made him do it. Alternatively, he could provide human reasons for being vegetarian, perhaps by talking about how it reduces gas emissions. For deliberative agency, the relevant answer to the why-question principally involves not so much getting answers to "how come?" questions but getting "what for"-answers or getting "reason explanations".

After all, when someone fires us, breaks up with us or tells us we should take a certain medicine, we want to know why, and we are usually not prepared to accept habits, brain activities or other causal stories as answers. What we usually want, especially also in digital contexts (medical or not), is reason-based explanation that enables us to understand why, say, you fit the profile that led to the decision, what kind of reasoning is underpinning the profile, why a clinician chose (not) to use AI tools in your case.

We want to stress that a 'useful' explanation here is understood functionally: what you need is the sort of knowledge that would put you in a position to evaluate and possibly *contest* the decision or course of action. Indeed, Ploug and Holm (2020) argue that explainability should be understood as "effective contestability". In order to ensure a patient-centric approach to AI diagnostics, they suggest that four different types of information should be provided: how the data in the AI system is used, the presence of possible biases, the system performance, and how the interaction between the system and health care professional is organized.

Of course, sometimes claims to receive knowledge of reasons may be ungrounded, and reason explanations might be poor or incomplete. But the basic idea that we can at least in some cases legitimately demand reasons of others is not just part of social etiquette or a nicety we grant one another, but is, we believe, morally significant – in life in general and thus often also in cases in which automated decision-making is involved.[12]

A brief comparison with lying

In the context of medical ethics, epistemic conditions of deliberative agency take on a very central role, given the fundamental role that informed consent plays in that context (see e.g. Beauchamp and Childress, 2013). Strangely, in the context of discussions about deliberative agency or autonomy there is much less attention to epistemic conditions. Typically, priority is given to discussions of various social and political conditions, such as the absence of oppression or manipulation, or the presence of self-trust or recognition.[13] There is one particular debate in ethics, though, that we believe is instructive for better understanding the moral importance of having a right to explanations.

A popular answer as to why lying is morally impermissible is because telling a lie involves taking an attitude towards another such that one fails to respect them as persons with a capacity to act on reasons of their own. When lying to someone, we are, according to P.F. Strawson's influential account, treating her not as "'a member of the moral community' but rather regard her as a *thing*: as something we can work around, nudge, control or manage" (Strawson, 1962, p. 18). If someone makes a 'lying promise' to another, for instance to repay some money, as in Kant's classic example, and the lender accepts, then the lender is consenting to the action 'giving my money away temporarily and having it returned later'. But the action, in reality, is giving the money away permanently, an action to which the lender certainly would *not* have consented. As Rae Langton (1992, p. 489) points out: "to deceive is thus to make a person thing-like: something that cannot choose what it does". In our terms, lying to someone is a clear violation of the epistemic conditions of deliberative agency.

We think Strawson's and Langton's general line of argument can be extended. In fact, Langton herself already paves the way for an interesting extension. She convincingly argues that what is morally problematic about lying is not limited to lying alone, it can also apply to simply "failing to tell the (whole) truth".[14] We welcome the extension

[12] Claiming that someone has a moral right to something is of course compatible with acknowledging that other (moral) considerations can sometimes trump that right. It is just to say that moral considerations have special weight and that they typically trump non-moral, prudential (e.g. commercial) reasons.

[13] Which is not to say that the focus on these other aspects is not legitimate. But it does raise the question of why epistemic conditions for deliberative agency, including in particular the right to knowledge of reasons, has not gotten the attention they deserve outside of medical ethics. A real possibility is that the presence of such epistemic conditions are so obvious that few find it necessary to explicate them. We agree it is obvious that one needs to know and understand things about the (social and digital) world in order to act, but we also believe it is not obvious at all how much and which kind of epistemic goods are necessary and when requests for explanation are legitimate and when they are not.

[14] It should be noted that Langton might well not agree with our argument here. Central to her account is whether a speech act is *strategic* rather than communicative and whether it counts as *deceptive* (1992, 490). She does not formulate the moral wrong of lying in terms of

and believe it can be broadened even further: we might think of explanations (that is, as we defined them, acquiring knowledge of reasons), too, as an important epistemic condition for deliberative agency, and we might think of not getting explanations in the relevant circumstances as violations thereof.

In certain cases, getting an explanation can, similar to being lied to, "make a person thing-like: something that cannot choose what it does". This would be the case if a patient were to get, upon requesting explanation, a highly technical explanation instead of a reasons-based explanation of how, say, a clinician who based her analysis on an AI-driven X-ray image came to recommend surgery or an explanation of why she chose to rely on the AI-system in this case. Leaving a patient in the dark after she requests an explanation, which importantly can also happen by giving her information that is meaningless *to her*, would strongly impair her deliberative agency because she would not be positioned to make important subsequent decisions (to go ahead, not to go ahead, to get a second opinion, and so on).[15]

Naturally, we do not have a general right to explanation – a right to know and understand reasons - for all of our why-questions. But we do plausibly have such a right, we want to propose, when the conditions mentioned above are met: when a) the knowledge in question is non-trivial and not getting an explanation inhibits or undermines one's core deliberative agency and b) there are other agents and/or institutions have access to, or are responsible for, providing the relevant epistemic goods.

We believe that the two conditions are fulfilled in many cases involving automated processing, especially in the medical domain. As to the first condition, it is generally accepted that health interventions are far from trivial and patients should be allowed to take up an active role in deciding on the course of action when it comes to their own health, and should understand the medical course of action they consent to, also if they involve the use of AI-systems. As for the second condition, recall from section 2 that it is never strictly speaking true that a machine 'made a decision' about you, even though it might seem this way. The duty-bearers in question are, for the time being, still human beings behind the algorithms and the machines. It's from *them* – human beings – that we want, and are owed, explanations. It is possible in theory that, at some point in the future, AI-systems (such as highly advanced and personalized 'medical dialogue bots') might well be able to give patients equally satisfactory, or perhaps even better, reason explanations than humans. Acknowledging this possibility is part of accepting the symmetry thesis. For the time being, however, this is science fiction, and the best explanations to receive about (medical) decisions involving automated processing will have to come from fellow humans – limited as they inevitably are.

Conclusion

We have argued in this paper that the solid normative intuition that we have a right to explanation when we are subject to algorithmic decision-making itself calls out for an explanation. Do we have such a right at all, and why? We have considered, and rejected, the 'asymmetry thesis' that grounds the normative right to explanation in what we have suggested is an implausible metaphysical view of 'human' versus 'non-human' decision-making. In its place we defended the 'symmetry thesis' according to which there is no special normative reason to have a right to explanation when 'machines' in the broad sense make decisions, recommend treatment, discover tumours, and so on. Instead, we argued that the specific right to explanation in contexts involving automated processing is derived from a general moral right to explanation when decisions are made that significantly affect us but which we do not (properly) understand. We have a right to know the reasons why a friend or our partner is late without our understanding why, just as we have a right to know why 'computer says no' when it comes to deciding for or against some medical procedure. The reason that dominantly automated decisions call out for explanations more acutely, or lead to stronger normative intuitions, is contingently due to the fact that we often – though not always – understand the way algorithms and

violations of the epistemic conditions of deliberative agency. Our strategy by contrast involves asking whether the absence of explanations can be morally problematic *even if* they are not deceptive in any straightforward sense. Alternatively, one might say that not acquiring knowledge of reasons when one has a right to (e.g. in automated-decision making processes) is itself a form of deception. We shall leave this discussion aside for now.

[15] One might wonder whether 'informed consent' and having a 'right to explanation' aren't in fact the same notions (thanks to a referee for raising this question). To clarify, the notion of having a 'right to explanation' as we discuss it in this paper is initially coupled specifically to automated decision-making, which informed consent is not. However, in this paper we broaden the notion of the narrowly defined AI-related right to explanation by showing how such a right in fact is grounded in a general moral right to explanation. Here, we thus arrive at a clear connection to informed consent, which likewise – or so many argue - gets its normative force from the fact that informed consent is necessary for deliberative agency (cf. Dworkin, 2010). However, whereas informed consent is something that is paradigmatically required before decisions are made and medical procedures are carried out, a right to explanation can take on both ex-ante and ex-post forms. Furthermore, consent is something given by the individual (or patient) and is an activity, whereas a right to explanation is just that: a right, and is thus something an individual might have even in contexts that do not involve her actively giving consent. In fact, even other parties or persons might 'stand up' for her right, even if she herself does not. In short, these notions are closely related but can and should nonetheless be distinguished.

medical AI systems work a lot less than our friends and partners. Even if they too often remain incomprehensible to us and can leave us in the dark, we typically at least understand the type or *shape* that the answers to our why-questions take.

References

Aagaard ,J., Friis, J. K. B., Sorenson, J., Tafdrup, O., Hasse, C., & Rosenberger, R. (eds) (2018), *Postphenomenological Methodologies. New ways in mediating techno-human relationships*. Edited by Rosenberger, R., Verbeek, P. P. & Ihde, D.. Lanham: Lexington Books

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE* access 6:52138–52160

Ananny, M., & Crawford, K. (2018) Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society*, 20(3):973–989. https://doi.org/10.1177/1461444816676645

Arpaly, N. (2002). *Unprincipled Virtue.* Oxford University Press, Oxford

Anscombe, G. E. M. (1957). Report on Analysis' Problem' no. 10. *Analysis*, 17(3), 49-53.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020). Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. *Operational AI Ethics*. https://hal.telecom-paris.fr/hal-02506409HAL Id: hal-02506409

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.

Buss, S., & Westlund, A. (2018). Personal Autonomy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 ed.). Meta-physics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Coons, C., & Weber, M. (Eds.). (2014). *Manipulation: theory and practice*. Oxford University Press.

Dennett, D. C. (1981). True believers : The intentional strategy and why it works. In A. F. Heath (ed.), *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*. Clarendon Press. pp. 150--167.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., Shieber, D., O'Brien, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *SSRN Electron J*. https://doi.org/10.2139/ssrn.3064761

Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Communicationsof the ACM* 59 (2): 56–62. https://doi.org/10.1145/2844110

Duncan, Pritchard (2009). Knowledge Understanding and Epistemic Value. *Royal Institute of Philosophy* Supplement 6419-43 10.1017/S1358246109000046

Eilan, N. (2014). The You Turn. *Philosophical Explorations* 17(3):265–278. https://doi.org/10.1080/13869795.2014.941910

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 *IEEE 5th International Conference on data scienceadvanced analytics (DSAA)*

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 *IEEE 5th International Conference on data scienceadvanced analytics (DSAA)*

Goodman, B., & Flaxman, S. (2016). European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" *ArXiv:1606.08813 [Cs, Stat]*, June. http://arxiv.org/abs/1606.08813

Goodyear-Smith, F., & Buetow, S. (2001). Power issues in the doctor-patient relationship. *Health Care Analysis* 9(4):449–462

Goodyear-Smith, F., & Buetow, S. (2001). Power issues in the doctor-patient relationship. *Health Care Analysis*, 9(4):449–462.

Hatherley, J. J. (2020). Limits of trust in medical AI. *J Med Ethics*, 46(7):478–481.

Hatherley, J. J. (2020). Limits of trust in medical AI. *J Med Ethics* 46(7):478–481

Heal, J. (2013). Social anti-individualism, co-cognitivism, and second person authority. *Mind*, 122(486), 339-371.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press, Bloomington

Kahneman, D. (2013). *Thinking, Fast and Slow*. Farrar, Straus and, New York Giroux

Korsgaard, C. M. (1983). Two Distinctions in Goodness. *Philos Rev* 92(2):169–195. https://doi.org/10.2307/2184924

Kvanvig, J. L. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.

Latour, B. (1992). Where are the missing masses. In *Shaping Technology/building Society: Studies in Sociotechnical Change*, edited by E. Bijker and J. Law, 225–258. Cambridge: MIT Press

Latour, B. (1993). *We have never been modern*. Harvard University Press, Harvard

Langton, R. (1992). Duty and Desolation. *Philosophy* 67(262):481–505

Latour, B. (1992). Where are the missing masses. In *Shaping Technology/building Society: Studies in Sociotechnical Change*, edited by E. Bijker and J. Law, 225–258. Cambridge: MIT Press

Latour, B. (1993). *We have never been modern.* Harvard University Press, Harvard.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep* 49(1):15–21. doi: https://doi.org/10.1002/hast.973

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1):15–21. doi: https://doi.org/10.1002/hast.973

Mackenzie, C., & Stoljar, N. (2000). *Relational Autonomy: Feminist*

Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *ArXiv:1706.07269 [Cs]*, June. http://arxiv.org/abs/1706.07269

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–88. https://doi.org/10.1145/3287560.3287574

Noorman, M. (2021). Responsibility and Liability. In *A citizen's guide to artificial intelligence*, edited by J. Zerilli, 61–79. Massachusetts: MIT Press

Perspectives on Autonomy, Agency, and the Social Self. Oxford University Press

Noorman, M. (2021). Responsibility and Liability. In *A citizen's guide to artificial intelligence*, edited by J. Zerilli, 61–79. Massachusetts: MIT Press

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics* 24(4):1201–1219. https://doi.org/10.1007/s11948-017-9943-x

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1 edition. New York: Crown

Oshana, M. (2014). *Personal Autonomy and Social Oppression: Philosophical Perspectives.* Taylor & Francis.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press

Ploug T, and Holm S (2020) The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artif Intell Med* 107:101901

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine,* 107:101901

Rosenberger, R., & Verbeek, P. P. (eds) (2015). *Postphenomenological investigations. Essays on human-technology relations.* Lexington Books, Lanham

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI* 5. https://doi.org/10.3389/frobt.2018.00015

Selbst, A., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law* 7(4) 233-242 10.1093/idpl/ipx022

Selbst, A., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *SSRN Electronic Journa*l 10.2139/ssrn.3126971

Strawson, P. F. (1962). Freedom and Resentment. In *Freedom and Resentment and Other Essays*, 1 edition. London; New York: Routledge

Verbeek, P. P. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences* 7(3):387–395. doi:https://doi.org/10.1007/s11097-008-9099-x

Wadden, J. J. (2021). Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics*

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent explainable and accountable AI for robotics. *Science Robotics* 2(6) eaan6080 10.1126/scirobotics.aan6080

Wellner, G., & Rothman, T. (2020). Feminist AI: Can we expect our AI systems to become feminist? *Philos Technol* 33(2):191–205