AI & SOCIETY (2025) 40:1695–1704
https://doi.org/10.1007/s00146-024-01939-7

OPEN FORUM

Is explainable AI responsible AI?

Isaac Taylor[1]

Received: 3 January 2024 / Accepted: 29 March 2024 / Published online: 20 April 2024

Abstract
When artificial intelligence (AI) is used to make high-stakes decisions, some worry that this will create a morally troubling responsibility gap—that is, a situation in which nobody is morally responsible for the actions and outcomes that result. Since the responsibility gap might be thought to result from individuals lacking knowledge of the future behavior of AI systems, it can be and has been suggested that deploying explainable artificial intelligence (XAI) techniques will help us to avoid it. These techniques provide humans with certain forms of understanding of the systems in question. In this paper, I consider whether existing XAI techniques can indeed close the responsibility gap. I identify a number of significant limits to their ability to do so. Ensuring that responsibility for AI-assisted outcomes is maintained may require using different techniques in different circumstances, and potentially also developing new techniques that can avoid each of the issues identified.

Keywords Artificial intelligence (AI) · Explainable artificial intelligence (XAI) · Moral reasoning · Responsibility gap

1 Introduction

Existing and possible future artificial intelligence (AI) systems may take high-stakes decisions out of the hands of humans. In many cases, the systems themselves may need to make choices that have ethical aspects. Self-driving vehicles are already on public roads and, as these increase in sophistication, the systems may need to be programmed to respond to ethical dilemmas in accident scenarios. Within a matter of years, it may be possible to develop lethal autonomous weapons systems (LAWS), which would be capable of selecting and engaging targets in an armed conflict without direct human control. Finally, risk-assessment software is currently being used in a number of jurisdictions to assist judges in making sentencing and parole decisions for convicted criminals. It might be possible to develop more sophisticated software that could offer a recommendation of what sentence should be given.

The movement toward algorithmic decision-making in these and other areas may have some benefit. AI systems are not subject to the same biases, irrationalities, and immoral tendencies as humans, and so theoretically could make better decisions on the whole. Nonetheless, some authors still worry about delegating too much power to these systems. One notable concern is the idea that using certain AI systems will create a morally problematic "responsibility gap": that is, a situation in which nobody is responsible for the actions taken by the AI system (Mathias 2004; Sparrow 2007).

As we will see, one significant potential source of responsibility gaps in AI development is the opaque nature of some of these systems. Because their inner-workings cannot be understood by even their creators, nobody can be properly held morally responsible for the systems' behavior, since they could not have reasonably predicted it. Nonetheless, recent advances in explainable artificial intelligence (XAI)—a cluster of techniques that are supposed to provide humans with greater knowledge of how autonomous systems function—might be and have been thought to provide a way out. According to one recent statement of this idea, '[o]nce a meaningful explanation of the recommendation is available to the decision-maker, we can more easily bridge the responsibility gap' (Baum et al. 2022: 14).

In this paper, I consider whether, and under what circumstances, existing XAI techniques can bridge responsibility gaps. While sufficient for some purposes, different forms of XAI are subject to different limits in this respect. In Sect. 2, I explain in greater detail why responsibility gaps might be thought to arise because of a lack of knowledge. In Sect. 3,

✉ Isaac Taylor
isaac.taylor@philosophy.su.se

[1] Department of Philosophy, Stockholm University, 10691 Stockholm, Sweden

I consider whether what we can call "feature-based" XAI techniques can close responsibility gaps of this sort, and in Sect. 4, whether "reason-based" XAI techniques can. I conclude that the success of closing the responsibility gap using XAI will crucially depend on *which* XAI techniques we deploy. Not all XAI techniques will be suitable in all cases.

2 When do responsibility gaps arise?

Not everyone thinks that there are troubling responsibility gaps when AI systems are deployed. In a challenge to those who posit the existence of such gaps, Peter Königs (2022) notes that two questions must be answered if such a defense is to be successful: First, when do responsibility gaps occur? And, second, why are they morally problematic? Answering the second question is beyond the scope of this paper, but we can note that authors have argued that there are moral costs attached to responsibility gaps, both instrumental (Danaher 2016; Sparrow 2007: 67; Taylor 2021: 322) and non-instrumental (Sparrow 2007: 67–68, 2016: 106–110). In this section, however, I provide a partial answer to the first question. My aim is not to provide a full defense of the existence of responsibility gaps, but rather to explicate one significant way in which they might arise. This will be important, because noting the circumstances under which there might be responsibility gaps will be important to see how XAI might be used to fill them.

Before getting to this task, some preliminary remarks about responsibility are needed. In discussions of responsibility gaps, the operative notion is that of *moral* responsibility, which links an agent with an action (or, sometimes, an outcome) in a morally significant way. This is different from merely *causal* responsibility, which is an account of how one action is caused by an agent or other entity without any ethical upshot. When someone is morally, and not merely causally, responsible for something it may, in contrast, make sense to blame or praise them for that thing.

It is generally thought that, for an agent to be morally responsible for an action, two conditions must be met (Fischer and Ravizza 1998: 12–14; Rudy-Hiller 2018). First, they must have control over the action (the "control condition"). And, second, they must have a degree of awareness about the action, including its consequences and perhaps broader moral significance (the "epistemic condition"). Someone who is forced to act in a certain way, or who is unaware of what they are doing, cannot be considered morally responsible.

Why might the use of AI systems undermine responsibility? Some have suggested that the control condition is undermined when these systems are used. Because of the large numbers of people who are involved in the life-cycle of AI systems (programmers, developers, users, and so on), it might be thought that no one individual's actions will have an effect on the overall outcome. However, one might wonder if this really does undermine individual responsibility (Oimann 2023: 9–10). Moreover, it might be thought that ascribing responsibility to the group as a whole would be sufficient to avoid the problems associated with a responsibility gap of this sort, at least in some cases (Conradie 2023; Taylor 2021).

In any case, my focus here will be on a second way in which responsibility gaps might arise. This is through the undermining of the epistemic condition. Some think, for example, that certain sorts of AI systems will act in ways that cannot be foreseen by any human. Due to the unpredictable behavior of these systems, no person in the life-cycle can reasonably be expected to know what they will do in new scenarios. Consequently, even the control condition is met here, the epistemic condition is not: the individuals in question do not know how to use their control to change AI systems' behavior in desirable ways. Moral responsibility is lacking in these cases.

Of course, not every deployment of an AI system will create a responsibility gap. In cases where AI systems undertake simple tasks, users may be able to accurately predict their behavior. If a self-driving car is programmed to maintain a constant speed, for instance, users may be morally responsible for failing to break for pedestrians, since they can foresee that the car will continue moving without such interventions. However, those who claim that responsibility gaps exist would generally hold that there are differences between the sort of simple algorithm guiding this self-driving car and more complex *autonomous* systems. Broadly speaking, we might understand an autonomous system as a system that is, in one or more respects, "self-governing".[1] And, it has been claimed, responsibility gaps are more likely to arise with respect to these sorts of systems. The rough idea would be that, since humans are not guiding central aspects of these systems, they cannot know what they will do, and thus cannot be responsible for the behavior that occurs. Robert Sparrow, in his influential application of this idea to LAWS, writes that 'the possibility that an autonomous system will make choices other than those predicted and encouraged by its programmers is inherent in the claim that it is autonomous' (Sparrow 2007: 70).

There are, in fact, two senses in which a system might be autonomous, both of which are relevant here. The first relates to how the system is programmed. In contrast to "top-down" systems where programming is completely written in

[1] The first instance of defining autonomous AI in terms of its self-governing nature that I am aware of is found in US Department of Defense (2012: 13–14).

advance (as is likely to be the case of the simple self-driving car in our example), "bottom-up" systems might write and re-write their own programming to achieve specified success conditions (Tasioulas 2019: 52). It is this capacity that the first sense of autonomy refers to. While the "higher-level" tasks of autonomous systems of this sort (such as arriving at a destination safely, if we are talking about a self-driving car) are specified in advance, the "lower-level" tasks necessary to complete the higher-level tasks (such as selecting a route and a speed) are more and more left to the discretion of the system as autonomy increases (Sartor and Omicini 2016: 42–43). And with this sort of autonomy, comes unpredictability. Although we know what the goal the system will be trying to achieve, we may not know what it will do to achieve it. Consequently, we may not hold those who develop or use AI systems responsible for their behavior as they become more autonomous.

As an example of how this can occur, we might consider the technical details of the most common form of autonomous system: a (supervised) artificial neural network. These systems mimic the functioning of the human brain to complete tasks when top-down programming is infeasible. They have been particularly successful at developing image-recognition technology. Suppose, for example, that we develop a LAWS that is supposed to identify tanks and fire on them once them does. It would, of course, be impossible to program a set of rules for identifying tanks—there would have to be a huge number of such rules given the different ways in which tanks could appear to the LAWS. So this system will need to decide how to complete the task. How would it go about this?

The LAWS might receive a visual input—a picture of the battlefield—via a camera. This picture would consist of n pixels, each which would be converted into a numerical representation. At its simplest, a dark-colored pixel might be assigned a number "0" and a light-colored pixel a number "1" (although a more complex system might assign a decimal to pixels whose color gradient is somewhere between black and white—0.1, 0.5, 0.75, etc.).

Pixels in isolation provide little information. What the system now needs to do is identify salient *patterns* among the pixels. If there is a long, metal-colored shape in the image, for instance, this might be a tank gun. If enough patterns like this are present, there is a good probability that there is a tank in the picture. To spot patterns, the system will use the values for various sets of neighboring pixels and combine them in a number of functions, each of which will give a set of new values between 0 and 1. The process might be repeated several times with each new set of values: various combinations of these will be transformed via a function into a new value. At the final stage, all the values produced at the previous stage will be combined into a single function, producing a value between 0 and 1 which will determine—with some degree of uncertainty—whether a tank is present in the picture or not. If this final value is above 0.5, say, it might be sufficiently certain that a tank is present and instruct the system to open fire. If not, no instruction to fire will be forthcoming.

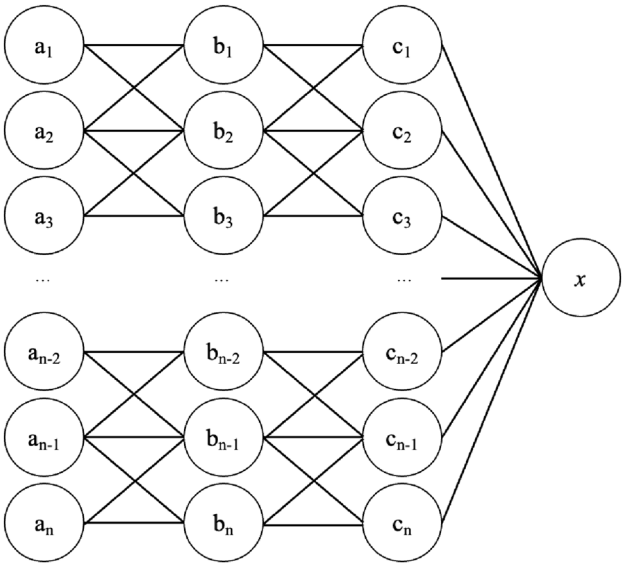

Fig. 1 An artificial neural network

This artificial neural network is represented visually in Fig. 1. The nodes in the column on the left represent the input data—each node represents a pixel, and the "a" values within them are the values assigned by the system based on the colour of each pixel. These nodes are ordered such that adjacent nodes represent adjacent pixels in the image: a_1 and a_2 represent pixels that are next to each other on the image, for example, and a_1 and a_n represent pixels that are at opposite corners of the image. The next layer of nodes—the column to the right of the input layer—contain "b" values, which are the values produced by the function of the "a" values from the connected nodes (e.g., $b_1 = f(a_1,a_2)$; $b_2 = f(a_1,a_2,a_3)$; etc.). There is a third layer producing "c" values—functions of the "b" values from connected nodes. Finally, all these "c" values are combined in a function that give an output value—x—which determines whether the LAWS will fire or not.

This system might be quite complex. There could be many pixels in an image; n could be in the thousands. Consequently, it would be infeasible for any human to program this from scratch. Rather, machine learning is needed—the system will need to work out for itself which patterns to look for. This can be done in a training phase where a variety of pictures, some including tanks and some not including tanks, are fed into the system, which learns which patterns are salient, and designs the neural network to effectively detect these. It might also be partially done once the system

is operational—the system can be provided with feedback by humans about its success rate in detecting tanks in the field and can continue refining the neural network on the basis of this to improve accuracy.

What this means, however, is that human understanding of how the system works might be limited. No human could get a grip on how the initial pixel values are transformed into an output value, given the vast number of nodes at each stage. This is why the layers of the network between the inputs and output are often referred to as "hidden": for all intents and purposes, they cannot be fruitfully examined by human beings. This is why it might be thought that individuals involved in the design and use of artificial neural networks fail to meet the epistemic condition: they are unable to fully understand how their actions (the initial programming and training) affect the eventual behavior of the system.

One might think that we could ensure responsibility for even autonomous systems by ensuring a "human in the loop", as it is often called. By this, it is meant that, although an autonomous system will arrive at a recommendation for an action, a human user must confirm this before the action is carried out. However, we sometimes not only have reasons for wanting systems to be autonomous in the first sense that I have outlined (because, for example, such systems will be able to perform better than humans in writing the programming that guides their behavior). We might also want autonomy in the sense of actually acting freely of human oversight and control. This is the second form of autonomy that I want to highlight: the capacity to act without the possibility of intervention by other agents. This type of autonomy might be valuable because having a human in the loop in some cases would be impossible (as when a self-driving car must take a split-second decision in an accident scenario). Or it might be because we do not want (or cannot guarantee) the means through which control can be exercised (such as through radio signals in a war, when these signals might be disrupted or when radio silence is essential for the success of the operation). Finally, we may simply think that the AI systems are making a decision, any human second-guessing that decision will result in the benefits of AI being lost. If a sentencing algorithm takes more information into account than any human could, it is unclear how a human could make a principled decision to accept or overrule that decision. In these cases, a human in the loop is an unavailable or undesirable response to the responsibility gap.

It may be questioned whether autonomy necessarily leads to unpredictability or, even if it does, whether this unpredictability really is a problem. On the first point, we might think that, with sufficient testing, we can gain some knowledge of what an AI system will do when deployed (cf. Köhler 2020: 3137–3138). But, in many cases, this will not be the case. The very fact that we want systems to be autonomous in the first sense is often that they will experience novel scenarios that require tailored responses (especially in war, for example). And, as the novelty of the scenarios encountered increases, so will the potential for behavior that differs from what can be expected on the basis of training data (and even previous deployments).

On the second point, we might think that even if autonomous systems act unpredictability, human designers and users can still be morally responsible for taking a risk that harms will occur. While this is true, we might also care about responsibility for each action taken by an autonomous system as well. Responsibility for this is not guaranteed even if we can assign moral responsibility for the risk-imposition (Taylor 2024). Of course, there are many harms that occur that nobody is responsible for even when AI is not used. How does lack of responsibility for AI systems' behavior differ from these genuine accidents? This is an important challenge to those who hold that responsibility gaps are a problem, and I lack the space to provide a full response here. All I can do is signal that I am prepared to accept that the lack of a responsible agents for harms may potentially be a problem in many different situations when harms occur. What is significant about AI is that its use will increase the number of such cases.

The question we must consider, then, is whether we can get the benefits of autonomous systems while maintaining moral responsibility for decisions. And, in particular, can the use of XAI get us the best of both worlds?

3 Feature-based XAI

While the motivation for using XAI is not usually to fill responsibility gaps,[2] at least some authors suggest its potential for doing so (Baum et al. 2022; Langer et al. 2021: 6–7). Moreover, since we have seen that responsibility gaps can arise from a lack of knowledge on the part of human users (a point also made in Santoni de Sio and Mecacci 2021: 1064–1065; Sparrow 2007: 70; Taddeo & Blanchard 2022: 6–8), XAI may seem like a natural solution to them. Since XAI techniques are designed to provide humans with knowledge of how opaque AI systems work, they may seem perfectly suited to the task of ensuring that the requisite knowledge is met for moral responsibility—i.e., ensuring that the epistemic condition is met. This and the following section assess XAI's prospects on this front.

We might broadly distinguish between what can be called "feature-based" approaches in XAI from "reason-based"

[2] According to the framework provided by the AI4People Scientific Committee, for instance, a principle of Explicability (which requires, in part, the intelligibility of AI systems) is supposed to ensure that AI is designed in a way that ensures benefits, avoids harms, maintains autonomy, and achieves justice (Floridi et al. 2018: 699–700).

approaches. Let us begin by considering the former. Feature-based approaches provide us with information about the role that inputs play in the algorithm of an AI system. One widely-used feature-based technique is Layerwise-Relevance Propagation (LRP). This provides users with an indication of which inputs contributed most strongly to a given output.[3] If we have a risk-assessment program being used in the criminal justice system, for example, using LRP might give an indication about which factors related to a given defendant (their sex, their criminal history, their age, and so on) had the greatest effect on the risk-assessment score that is produced. LRP has also been used with visual recognition technology. In our imagined example of a LAWS in Sect. 2, it might tell us which pixels were most significant in characterizing a particular object as a tank, for example. This could be represented visually as a heat map: a copy of the input image with the most significant pixels highlighted (Binder et al. 2016).

How might feature-based techniques like LRP be thought to ensure that humans meet the epistemic condition on moral responsibility? By knowing which features were the most significant, it might be suggested, we can gain enough knowledge about how the system is operating to be able to predict its behavior. Consider the following famous (and possibly apocryphal) story. A nation's armed forces were attempting to build an artificial neural network that could recognize tanks visually (in the way described above). During the training phase, the system was fed a number of different pictures—some with tanks in, some without—and it developed a rule for identifying images of tanks. The rule, however, took the following form: a "tank" was taken to be "something with a forest behind it" by the system. This is because all the training images containing tanks had a forest in the background.

Without XAI, this discovery may not have been possible. The system may ultimately have been incorporated into weapons systems, with disastrous results. Perhaps we could not have expected individuals who designed it to know any better in that case—the system may have performed perfectly in tests. But, once LRP was used, they would have been able to see that the pixels that were most significant when tanks were identified in images were not the ones that made up the representation of the tank itself, but rather the pixels making up the scenery behind the tank. If the creators and users of this system decided to deploy it in a warzone anyway, we could properly hold them morally responsible for the behavior of the system for the resulting undesired outcomes, and blame them for the deaths that resulted. LRP would appear to ensure that the epistemic condition is met.

However, in other cases, it looks like LRP—and feature-based approaches to XAI more generally—will fail to bridge the epistemic gap. They key to understanding why is to begin by noting that the sorts of systems we are concerned with will not only be required to form reliable "beliefs" (for want of a better term) about their surroundings (determining that the object ahead is a tank and not an ambulance, for example). They will also need to engage in *moral reasoning* to know how to act in response to these beliefs.

We noted previously that many cases where a responsibility gap is thought to arise involves the use of AI to make decisions that have an ethical element. Indeed, the issue of moral responsibility might be thought to only arise with respect to ethically-laden choices. Consequently, these systems would need to be responsive to ethical reasons if their use is to be acceptable: at the very least, the systems must be sensitive to certain ethical constraints. Now these constraints might be incorporated into the systems in one of two ways. First, the systems might contain rules that directly constrain their behavior from the outset. Some authors argue that these sorts of rules could be programmed into the systems in a top-down manner (Arkin 2009: 99–104; Zając, 2020). But while this top-down approach might be sufficient to ensure desirable outcomes on some occasions it is thought that more than this may be needed to ensure the ethical behavior of robots. It has been argued that autonomous systems should be designed to act as (or like) virtuous agents—determining right action from considering the morally relevant features of a situation—rather than simple rule-followers (Abney 2013: 347; Wallach and Vallor 2020). This is the second way in which constraints might be operative in autonomous systems.

Take, for example, what would be needed for LAWS to act in an ethical manner. This would require that the relevant ethical rules of armed conflict are followed before making the decision to engage a target. One of these rules is proportionality. On the standard understanding, it requires the military advantage gained in any military operation in a war not be morally outweighed by the unintended collateral damage done to civilian populations in the process. Now it seems unlikely that rules like the proportionality requirement are the sort of thing that could be programmed. This deceptively simple principle is in fact difficult to specify with sufficient precision. How, for example, are we to program a way of determining the value of military advantage and loss of life on a single scale? Noel Sharkey notes that 'the phrase 'excessive in relation to the concrete and direct military advantage expected to be gained' is not a specification. It is also practically impossible to calculate a value for 'the actual military advantage'' (Sharkey 2010: 380; cf. Sharkey 2012: 790).

In order for LAWS to act in line with principles like proportionality, then, they must be built with a capacity of moral sensitivity to the scenarios they encounter. This might be done through artificial neural networks as well—the

[3] For more detail, see Montavon et al. (2019).

inputs in this case would be morally salient facts about a situation that is confronting the system. If LAWS incorporated this model, the inputs might include, for example, the number of humans who are within a target area, their (non-) combatant status, and the strategic importance of the objective that a potential attack would accomplish—all expressed as a numerical value between 0 and 1. These would then be transformed through a series of functions to provide an output which determined whether a possible action was ethical or not. How could such a network be developed? Perhaps it could be fed training scenarios—some of which are designated as legitimate engagements and some of which are designated as unethical. The system would then learn to distinguish one from the other so that it can make well-founded ethical distinctions on the battlefield. This process might be thought to mirror the way in which Aristotle thought that humans can learn to be virtuous through taking note of how virtuous people—moral exemplars—act in various circumstances (Aristotle 1984: 1094b11–27; on the analogy with machine learning, see Wallach and Vallor 2020: 395).

What all this means is that, if the responsibility gap is to be filled by feature-based XAI, these techniques must give human agents sufficient knowledge of the moral reasoning behind a decision taken by an autonomous system, and not simply the ways in which the systems can gain information or the top-down rules that are programmed. Only then will they have enough knowledge to be able to predict how the systems will behave, and change their own actions accordingly to bring about better outcomes when necessary. It is my contention, however, that feature-based XAI will often be unable to do this.

How exactly does good moral reasoning proceed? In asking this, we might be understood as asking how individual moral reasons give rise to an all-things-considered reason for action. Some ethicists are holists about moral reasons: they believe that 'a feature that is a reason in one case may be no reason at all, or an opposite reason, in another' (Dancy 2004: 7). Reasons, on this view, can combine in quite complex ways. What someone who wants to understand why a certain decision was made needs is not merely to know which reasons had the largest impact, but how various reasons interacted with each other in a process of deliberation. However, feature-based XAI, as we have seen, can at most give us an indication of which reasons were in play. It cannot tell us how these reasons combine to produce behavior in autonomous AI systems.

As an example, consider a self-driving car that is programmed to minimize moral disvalue in accident scenarios. While driving at a high speed, it suddenly detects an individual directly ahead in the middle of the road, and must make a choice between allowing this person to be run over and swerving onto a side road and into another pedestrian. According to a plausible view, questions of responsibility should enter into how the car should behave in this sort of scenario (Kauppinen 2020). If the pedestrian on the side road is morally responsible for their being there (perhaps they ran into the road negligently but in full knowledge of the dangers of doing so), but the pedestrian on the main road is not, we might think that the ethical decision here is to swerve and hit the former. The fact that they have made themselves *liable* to be hit is a reason in favor of their taking on the unavoidable cost, rather than the "innocent" individual ahead.

Compare this with a second scenario of the same structure, except this time the pedestrian on the side road ran onto it to save a helpless child from danger. It may still be the case that this individual is morally responsible for being there—they acted freely, in full knowledge of the situation, and so on, in running into the road. Nonetheless, we might now think, their morally admirable motive nullifies what would otherwise be a reason for swerving into them. This motive, we might say, acts as a *defeater* (Bagnoli 2016): it makes what would otherwise be a reason for action (i.e., the moral responsibility of the pedestrian for their predicament) no reason at all when present.[4]

Knowing whether a self-driving car will act ethically, then, requires not only that we know what factors it is taking as reasons (the moral responsibility of potential victims, the moral value of their motives, and so on), but also how these factors are combined. And this combination can only happen in the hidden layers of an artificial neural network, which feature-based XAI cannot provide us with information about. When autonomous systems operate in scenarios where the appropriate ethical reasoning is sufficiently holistic, then, feature-based XAI cannot furnish us with the necessary understanding of how they will act.

Of course, the holist view of moral reasons is controversial. Perhaps we might think that morality is simpler than that. Utilitarians, for example, think that there is only one fundamental moral reason for action: the fact that a possible action would promote the welfare of sentient beings. If we wanted an AI system to act in line with this, could we be more hopeful of feature-based techniques providing us with the sort of knowledge that would render us morally responsible for the actions of the systems in a greater range of cases?

I suggest not. The problem here is not the messiness of morality, but the messiness of reality that a utilitarian system would need to confront when applying its principles. The

[4] Could a self-driving car really detect all the information needed to reason like this? Kauppinen (2020: 640) is optimistic that future AI systems could pick up on subtle and reliable cues about moral responsibility, and there is no reason to think that they could not also use heuristics to make judgements about motives. (Does the individual in the road have a child in their arms? Are they wearing a paramedic uniform?).

simple rule of only firing on tanks (in our example earlier) would not be sufficient to ensure that utility is maximized by LAWS. What about other sorts of vehicles that are of military use? What if combatants in the opposing army (perhaps in an attempt to exploit the limits of LAWS) use decommissioned ambulances as military vehicles? What if neutralizing tanks would lead to large collateral civilian damage? To be able to take into account all of these issues, LAWS would need to acquire knowledge about different features of their environment, and reason about these to determine what the optimal action at any given time is. Simply knowing which factors are featuring in its decision may not be enough to properly understand a LAWS' behavior. Again, feature-based XAI will come up short.

While feature-based XAI may provide enough information to bridge responsibility gaps some cases, then, when the AI systems in question need to engage in complex moral reasoning, this may not be the case. Given that responsibility gaps appear most likely to occur when *autonomous* systems are making ethically significant choices, this may be a common problem, since one of the motivations for deploying these sorts of systems is that they can respond to ethically complex scenarios.

4 Reason-based XAI

Because feature-based XAI cannot give a sufficient account of the reasons why autonomous systems acted in particular ways, we might think that using to a form of XAI that provides an account of this will be more fruitful (Baum et al. 2022). "Reason-based XAI" is the term that I will use to designate these techniques.

The first thing to note about reason-based XAI is that it will not give us reasons that fully explain why a system produces outputs in all cases. The function of reason-based XAI is to give us some reasons about how a system functions (or might function) without making the nature of the system fully transparent (which is taken to be impossible). Three common reason-based XAI techniques are: providing examples (giving an explanation of why a system produced a single output); providing approximations (explicating a model that approximates but simplifies the workings of the system); and providing counterfactuals (explaining outputs by elucidating how changing certain inputs would change those outputs) (Speith 2022: 2241–2242).

There are limits to the ability of all these types of reason-based XAI to close responsibility gaps. Take, first, the technique of providing examples. When such techniques are used, the reasons behind an individual decision are given. Assuming the adequacy of such explanations, this might indeed be enough for responsibility in some cases. If, for example, a judge is given a sufficient explanation of why a sentencing algorithm is recommending a certain sentence, they would appear to be morally responsible for accepting or altering the recommendation. This is because they would have the right sort of knowledge to understand whether the recommendation was appropriate.

Issues arise, however, if we are dealing with systems that are also autonomous in the second of the two ways previously discussed. LAWS, as we have seen, would be autonomous not only in the sense of employing machine learning, but also in the sense of acting independently of human oversight and control. If humans are provided with explanations of why these systems acted as they did on specific occasions (why a certain target was selected, for example), this can only be after the fact. Unlike the case of the sentencing algorithm, no intervention can be made to alter behavior. Consequently, we cannot consider any humans as morally responsible for the action. Of course, with the knowledge provided by the example (and, indeed, in training scenarios), developers might adjust the system to improve future outcomes. But, as I noted earlier, autonomous systems often encounter novel scenarios (especially in battlefield settings), and so any knowledge of how it operated in previous scenarios will be of limited use.

Turn now to the second of the three types of reason-based XAI: the provision of approximations. Cynthia Rudin argues that this technique is problematic because of the failure of approximations to be faithful to how the original model operates. An approximation of a sentencing algorithm, for example, might suggest to us that a sentencing recommendation was made on the basis of race when, in fact, it was made on the bases of distinct factors that are nonetheless correlated with race (Rudin 2019: 207–208). Why might this matter? In some cases, simply knowing how a system recommends a distribution of burdens across social-salient groups might be sufficient to know when the distribution is morally problematic. Recent proposals for criteria of algorithmic fairness, for instance, are only sensitive to distributions in this way (Eva 2022; Hedden 2021). But that might not be all we care about. We might, for instance, want the treatment of individuals to not only (fail to) be correlated with certain factors, but also to (fail to) be actually made *because* certain factors are present. Some think, for example, that it is impermissible to allow factors that are beyond a defendant's control determine their punishment (Husak 2007). On this view, we would need to know which factors are, in fact, playing a role in setting punishment. Relying on approximations may thus fail to give us sufficient knowledge to determine whether a sentencing recommendation is appropriate.

How about the third type of reason-based XAI: the provision of counterfactuals? Suppose, for example, that a counterfactual explanation is provided with respect to an individual recommendation from a sentencing algorithm. This explanation would tell the user what would need to be

different about the inputs for the recommendation to change in a certain way. Such explanations might well provide the users with an indication of whether the algorithm is making recommendations on the basis of illegitimate reasons—if, for instance, it was shown that a lower sentence would be recommended if the defendant's race was different, the user can be fairly sure that the algorithm is not tracking legitimate reasons for longer sentences. If a user in such a scenario were to continue to use the algorithm, they would be blameworthy for the racially-biased sentencing that resulted. Nonetheless, counterfactual explanations, as I will now suggest, cannot track all relevant reasons.

This is because some moral properties that we might want AI systems to take account of are in some sense *comparative*. Such a moral property's magnitude will depend on the actions or treatment of others. While there may be a number of comparative moral properties of this sort (McLeod 2003), I will focus on one illustrative case.

Desert is sometimes to be taken to be a comparative moral property (Miller 2003). Whether a given runner deserves an Olympic medal will depend on how their performance rates against the other competitors, for instance. Something similar may be thought about negative desert, which is often taken to be an important consideration in determining criminal punishments. When we are talking about desert for punishment, of course, non-comparative elements are surely also relevant: punishing a mass murderer with a short, suspended prison sentence is surely not to give them what they deserve. But within the outer bounds set by these non-comparative elements, we might think that defendants desert is dependent on the punishments that are given out to others. We might say, for instance, that if one person receives a prison sentence 20 years higher than everyone else who committed the same crime (and there is no morally significant difference between them), the heavy punishment is undeserved.

We are now in a position to see why counterfactual explanations might be inadequate to meet the epistemic condition when comparative notions like desert are in play. Suppose that a sentencing algorithm is used to decide what length of punishment to give to defendants. While counterfactual explanations might be sufficient for judges to know in *some* cases whether they should overrule the algorithm (when race is a determining factor, as in our example above), they will not be able to know whether or not desert has properly been taken into account. This is because they would also need to know whether the proposed sentence was in line with the general sentencing practice in the jurisdiction.

Of course, while the general level of severity of sentencing in a jurisdiction could in theory be a variable input value, and thus feature in a counterfactual explanation, the determination of this variable would itself require complex calculations. Which other cases are relevant to take into account? And how should these be taken into account? These choices will themselves require ethical reasoning, and thus if users are going to be morally responsible for judgements made on the basis of them, they will need some way of knowing how moral reasoning proceeds. The problem of providing the requisite knowledge to humans will thus simply get pushed to a different site in these sorts of cases.

The lesson from this section is that, while reason-based XAI techniques perform better that feature-based techniques at ensuring the epistemic condition is met, each of the major forms of reason-based technique will be of limited use in certain sorts of cases.

5 Conclusion

We have considered whether deploying existing XAI techniques can bridge the responsibility gap. Despite what might be suggested by proponents of this idea—who seem to assume that (certain forms of) XAI might be a catch-all solution whenever the epistemic condition is undermined—we have found that different forms of XAI are subject to different limits on their usefulness. This suggests that care needs to be taken when deciding which form of XAI is used.

Feature-based XAI will be of limited use: it will fail to bridge the responsibility gap whenever the AI systems in question need to engage in complex ethical reasoning. Reason-based XAI techniques may appear to fare better, although different forms of reason-based XAI will be subject to different limits in different sorts of cases. The provision of examples will not help when we are dealing with systems that have a high degree of autonomy. Approximations run into problems when the ethical status of a decision depends on how it came about. Finally, counterfactual explanations will fail to provide adequate understanding of decision-making when they need to account for comparative moral properties like desert.

We might hope that choosing different sorts of reason-based XAI techniques depending on circumstances will ensure that the epistemic condition is always met. However, this assumes that the factors that impose limits on each of these techniques never appear together, and I do not think that this assumption can be taken for granted, especially if AI systems are used to make complex, multi-faceted decisions. Existing XAI may be no silver bullet capable of bridging every problematic responsibility gap. This said, if new reason-based XAI techniques can be developed that avoid all of the issues identified, we may have a more promising way of ensuring responsibility in a greater range of cases.

Acknowledgements I am very grateful to participants at the early career political philosophy workshop at Stockholm University for a fruitful discussion of this paper, as well as the reviewers and editors from this journal..

Funding Open access funding provided by Stockholm University. No funding was received to support this research.

Data availability Not applicable.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Abney K (2013) Autonomous robots and the future of just war theory. In: Allhoff F, Evans NG, Henschke A (eds) Routledge handbook of ethics and war. Routledge, Abingdon & New York, pp 338–351

Aristotle (1984) Nicomachean ethics, In: Barnes J (ed) The complete works of Aristotle: revised oxford translation. Princeton University Press, Princeton

Arkin RC (2009) Governing lethal behaviour in autonomous robots. CRC Press, Boca Raton

Bagnoli C (2016) Defeaters and practical knowledge. Synthese 195:2855–2875. https://doi.org/10.1007/s11229-016-1095-z

Baum K, Mantel S, Schmidt E, Speith T (2022) From responsibility to reason-giving explainable artificial intelligence. Philos Technol 35(1):1–30. https://doi.org/10.1007/s13347-022-00510-w

Binder A, Bach S, Montavon G, Müller K, Samek W (2016) Layer-wise relevance propagation for deep neural network architectures. In: Kim KJ, Joukov N (eds) Information science and applications (ICISA) 2016, Springer, pp 913-922

Conradie NH (2023) Autonomous military systems: collective responsibility and distributed burdens. Ethics Inf Technol 25:1–14. https://doi.org/10.1007/s10676-023-09696-9

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18(4):299–309. https://doi.org/10.1007/s10676-016-9403-3

Dancy J (2004) Ethics without principles. Oxford University Press, Oxford

Eva B (2022) Algorithmic fairness and base rate tracking. Philos Public Aff 50(2):239–266. https://doi.org/10.1111/papa.12211

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi SB, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Fischer JM, & Ravizza M (1998) Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press

Hedden B (2021) On statistical criteria of algorithmic fairness. Philos Public Aff 49(2):209–231. https://doi.org/10.1111/papa.12189

Husak D (2007) Rethinking the act requirement. Cardozo Law Rev 28:2437–2460

Kauppinen A (2020) Who should bear the risk when self-driving vehicles crash? J Appl Philos 38(4):630–645. https://doi.org/10.1111/japp.12490

Köhler S (2020) Instrumental robots. Sci Eng Ethics 26:3121–3141. https://doi.org/10.1007/s11948-020-00259-5

Königs P (2022) Artificial intelligence and responsibility gaps: what is the problem? Ethics Inf Technol 24:1–11. https://doi.org/10.1007/10.1007/s10676-022-09643-0

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (XAI)?—a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell 296:1–24. https://doi.org/10.1016/j.artint.2021.103473

Mathias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183. https://doi.org/10.1007/s10676-004-3422-1

McLeod O (2003) Comparative justice. In: Olsaretti S (ed) Desert and justice. Oxford University Press, Oxford, pp 123–144

Miller D (2003) Comparative and noncomparative desert. In: Olsaretti S (ed) Desert and justice. Oxford University Press, Oxford, pp 25–44

Montavon G, Binder A, Lapuschkin S, Samek W, Müller K (2019) Layer-wise relevance propagation: an overview. In: Samek W et al (eds) Explainable AI: interpreting, explaining and visualizing deep learning. Springer, pp 193–209

Oimann A-K (2023) The responsibility gap and LAWS: a critical mapping of the debate. Philos Technol 36:1–22. https://doi.org/10.1007/s13347-022-00602-7

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Learn 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudy-Hiller F (2018) The epistemic condition for moral responsibility. Stanford Encyclopedia of Philosophy

Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 34(4):1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Sartor G, Omicini A (2016) The autonomy of technological systems and responsibilities for their use. In: Bhuta N, Beck S, Geib R, Liu H-Y, Kreb C (eds) Autonomous weapons systems: law, ethics, policy. Cambridge University Press, Cambridge, pp 39–74

Sharkey N (2010) Saying "no!" to lethal autonomous targeting. J Mil Ethics 9(4):369–383. https://doi.org/10.1080/15027570.2010.537903

Sharkey N (2012) The evitability of autonomous robot warfare. Int Rev Red Cross 94(886):787–799. https://doi.org/10.1017/S1816383112000732

Sparrow R (2007) Killer robots. J Appl Philos 24(1):62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow R (2016) Robots and respect: assessing the case against autonomous weapon systems. Ethics Int Aff 30(1):93–116. https://doi.org/10.1017/S0892679415000647

Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Charles I, Lazar S, Oh A, Xiang A (eds) 5th ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, pp 2239–2250

Taddeo M, Blanchard A (2022) Accepting moral responsibility for the actions of autonomous weapons systems—a moral

gambit. Philos Technol 35(3):1–24. https://doi.org/10.1007/s13347-022-00571-x
Tasioulas J (2019) First steps towards an ethics of robotics and artificial intelligence. J Pract Ethics 7(1):61–95
Taylor I (2021) Who is responsible for killer robots? Autonomous weapons, group agency, and the military-industrial complex. J Appl Philos 38(2):320–334. https://doi.org/10.1111/japp.12469
Taylor I (2024) Responsibility for what? Reply to wood. Philos Technol 37:36. https://doi.org/10.1007/s13347-024-00729-9
US Department of Defense (2012) Autonomy in weapon systems. Directive 3000.09
Wallach W, Vallor S (2020) Moral machines: from value alignment to embodied virtue. In: Liao SM (ed) Ethics of artificial intelligence. Oxford University Press, Oxford, pp 383–412
Zając M (2020) Punishing robots—way out of sparrow's responsibility attribution problem. J Mil Ethics 19(4):285–291. https://doi.org/10.1080/15027570.2020.1865455