AI and Ethics
https://doi.org/10.1007/s43681-022-00172-6

ORIGINAL RESEARCH

Responsibility gaps and the reactive attitudes

Fabio Tollon[1,2,3]

Received: 20 January 2022 / Accepted: 6 May 2022

Abstract
Artificial Intelligence (AI) systems are ubiquitous. From social media timelines, video recommendations on YouTube, and the kinds of adverts we see online, AI, in a very real sense, filters the world we see. More than that, AI is being embedded in agent-like systems, which might prompt certain reactions from users. Specifically, we might find ourselves feeling frustrated if these systems do not meet our expectations. In normal situations, this might be fine, but with the ever increasing sophistication of AI-systems, this might become a problem. While it seems unproblematic to realize that being angry at your car for breaking down is unfitting, can the same be said for AI-systems? In this paper, therefore, I will investigate the so-called "reactive attitudes", and their important link to our responsibility practices. I then show how within this framework there exist exemption and excuse conditions, and test whether our adopting the "objective attitude" toward agential AI is justified. I argue that such an attitude is appropriate in the context of three distinct senses of responsibility (answerability, attributability, and accountability), and that, therefore, AI-systems do not undermine our responsibility ascriptions.

Keywords Reactive attitudes · Responsibility gaps · Artificial intelligence

1 Responsibility gaps and the reactive attitudes

In this paper, I would like to address the *fittingness* of our reactive attitudes, as these apply to agent-like AI systems. My concern in this paper, therefore, is to see whether individual preferences, in the form of the reactive attitudes, can be met with respect to our responsibility practices. What this entails is an investigation into the experiences that agents may have with technological systems, and whether certain demands made on artificial systems are *fitting* or not.

There has been an explosion of work in recently on the topic of responsibility gaps as these relate to AI systems [3, 8, 14, 15, 17, 24, 26, 31]. Most of this work has focused on whether a responsibility gap might emerge due to special properties that AI systems might come to possess which pose a threat to our responsibility practices. In this paper, however, my focus will be concerned with whether AI systems might undermine responsibility in a broader, social and psychological, sense.

There are two reasons that motivate such an investigation. First, due to the fact that these sentiments are experienced by agents and not some objective property of the world, it behooves us to take seriously the way AI systems might influence this experience. Second, AI systems are becoming increasingly ubiquitous. From social media timelines, video recommendations on YouTube, and the kinds of adverts we see online, AI, in a very real sense, filters the world we see [28]. Whether that world is rose tinted or stained red ultimately depends on what the algorithm thinks will maximise revenue for its owner in the best possible way. An important offshoot of this is that people will develop certain expectations from these systems, and may feel frustrated if these expectations are not met. Thus, I will investigate the so-called "reactive attitudes", and their important link to our responsibility practices. I will then show how within this framework there exist *exemption* and *excuse* conditions, and test whether our adopting the "objective attitude" toward agential AI is justified [25].

✉ Fabio Tollon
fabiotollon@gmail.com

[1] Department of Philosophy, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany

[2] Centre for Artificial Intelligence Research (CAIR), University of Pretoria, Pretoria, South Africa

[3] Data and Computational Ethics Research Group, Stellenbosch University, Stellenbosch, South Africa

Published online: 30 May 2022

1.1 Justification

Trying to understand *human* intelligence has been a preoccupation of our species that has proven rather fruitful. The more we understand how our minds work, the better we can treat, prevent, and cure disease and illness. Not only this, but there is the promise that might gain insight into the causal workings of our minds, which sheds further light on *why* we behave the way we do. Studying human intelligence involves analysing how it is that we perceive, predict, manipulate and understand the world around us ([20]: 1).

Of course, it would be well beyond the scope of this paper to say much more about human intelligence here, but it is worth keeping in mind as we canvass questions related to *artificial* intelligence. *Artificial* intelligence (AI) is "a cross-disciplinary approach to understanding, modelling, and replicating intelligence and cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices" ([10]: 1). AI attempts not just to *understand* intelligent systems, but also to *build* and *design* them ([20]: 1). Thus, this paper can be seen as part of the project of attempting to understand what happens when AI-systems, designed in specific ways, are embedded in contexts that, under normal circumstances with human agents, would demand responsibility-responses.

An important component of AI ethics that I leave out, but which of course is incredibly important, has to do with how the data that drives these systems is created, curated, and deployed. Human beings are essential in this process, and they are the ones who make the morally important decisions of how to label datasets, what a 'valid' dataset looks like, and what counts as a proper use of the technology in question. The harms of this can be widespread: from gender bias in hiring, racial bias in creditworthiness and facial recognition software, and sexual bias in identifying a person's sexual orientation, we are awash with cases of AI systematically enhancing rather than reducing structural inequality (Buolamwini and Gebru [2]; Gebru [11]; Nyholm and Gorda [18]). For the purposes of this paper, however, I must restrict my analysis to AI-systems that are *already embedded* in various contexts, and from that perspective, evaluate how users might respond to them. This is not to ignore the important issues that I have raised above but is instead a pragmatic recognition that these systems are already out there, and so it makes sense to try and get a handle on their potential moral consequences.

When an AI system performs an action that results in some event that has moral significance (and where we would normally deem it appropriate to attribute moral responsibility to human agents) it seems natural that people would still have emotional responses in these situations. This is especially true if the AI is perceived as having agential characteristics, which might come about if the system is capable of, for example, interacting with its environment without human control. If a self-driving car harms a human being, it would be quite natural for bystanders to feel anger at the cause of the harm (at least initially, for example, before they discovered that it was a self-driving car). Bystanders might incorrectly direct their anger towards the driver of the vehicle, who of course in this example does not exist. Upon discovering this, and assuming that nobody thinks the car itself is a fitting target of anger, we are left with a worrying situation in which it seems our anger has no fitting target. While such a situation might not create a *gap* in responsibility, it might nevertheless leave those involved feeling unsatisfied at not having their preferences or expectations met. This has implications for the overall functionality of society, and so is worth investigating in further detail.

The question then becomes whether this anger is *fitting* or *appropriate* (when addressed towards the AI) given the nature of the entity that was the cause of the harm. For an emotion to be fitting is "to think there is a (pro tanto) reason, of a distinctive sort, for feeling the emotion toward it" ([7]: 108). Essentially, we are interested in whether there is a reason for feeling a certain way. When we say a response is fitting, we are in a sense *endorsing* the response in that situation ([6]: 747). Thus, for a reactive attitude to be fitting, it should be the case that we have "reasons to feel" ([7]: 116). However, because emotions have motivational tendencies, attempting to regulate them is only an indirect way to influence behaviour ([7]: 111). This means that simply because we have determined that a given emotion (e.g., anger) was unfitting, does not necessarily mean that the agent will suddenly stop feeling it.

Many experts interested in the question of responsibility start their investigation by granting that we ought to (in the future, at least) extend some form of moral consideration to AI-based systems, and that this possibility motivates an investigation into responsibility [1, 4, 5, 12, 19, 21, 24]. However, I aim to bracket this question and instead argue, in a similar way to Coeckelbergh [5], that our *responses* to these systems matter, morally speaking. In my case, however, I do not mobilize an argument in favour of moral consideration of robots or AI, but am instead motivated by more pragmatic questions concerning the overall functioning of society.

My concern in this paper is not with whether there is a responsibility gap with respect to AI in a strict, or narrow, philosophical sense. Rather, the question I want to consider here is how an AI-based system might justifiably be exempted from responsibility altogether. That is, I want to investigate responsibility in a *broader* sense, which might not necessarily turn on a conceptual analysis of responsibility, but depends importantly on our natural sociality as a species. The question then becomes *why* our reactive attitudes are unfitting when directed towards AI systems. What

would constitute a justified suspension of our responsibility practices to that *specific* (AI-based) system? Having clarity on this matter will provide further evidence that our responsibility practices are not undermined by agential AI. Next, I will delve into some more specifics with respect to the Strawsonian account of moral responsibility.

2 Strawson and the reactive attitudes

In his influential *Freedom and Resentment* (1962) Strawson makes the compelling case that "traditional" ways of going about justifying our responsibility practices get things exactly backwards. Instead of looking toward objective or external conditions that need to be met in order for people to *be* morally responsible, we should look inwards, towards our actual practices of *holding* one another responsible. This 'looking inwards', however, should not be confused with introspection. That is, I do not mean to suggest that we ground moral responsibility in introspective states, but rather that we look towards how our emotions come to figure in our social practices. Thus the distinction between internal and external, as I use the terms here, simply refers to the direction in which we ought to look for our grounding of moral responsibility. That is, 'internal' refers to a pragmatic metaphysics, whereas 'external' refers to a more structured, conceptually dense, metaphysics. For now, back to Strawson.

When we understand responsibility in this way, according to Strawson, we find that our "reactive attitudes" (such as resentment) are constitutive of these practices, and not merely inconsequential aftereffects [33]. This pragmatic approach sidelines discussion of free will or whether our world is deterministic by accounting for all of our responsibility responses in terms of the reactive attitudes.

Strawson begins his analysis by focusing on a specific class of emotions—what he calls the reactive attitudes, which he argues play a constitutive role in the way we hold one another morally responsible. Strawson believes that these emotions are the bedrock upon which our practice of holding responsible rest, and that this is due to a natural disposition shared by all human beings to care about what others think of them. In Strawson's own words, this comes about because of the

> "very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions (1962: 3)"

When we add interactive technology to this equation something interesting happens. Namely, the possibility is raised that our personal feelings and reactions come to be attached not just to human beings but also to artificial systems. This is not to say that these systems have the same *moral status* as human beings, but that humans, in their interactions with these systems, might still come to *expect* certain things from them, and should these desires not be fulfilled, might be disappointed or confused at such an outcome. And sure enough, people already (a) come to form attachments to technological devices, and (b) these devices are being designed in ways that explicitly promote such attachment. Think of PARO, the therapeutic and interactive robot, designed to provide comfort to those in nursing homes or hospitals.[1] Patients in the presence of such robots are encouraged to develop attachments to them, and this might lead to certain expectations. For example, consider this quote from the PARO website: "By interaction with people, PARO responds as if it is alive, moving its head and legs, making sounds, and showing your preferred behaviour. PARO also imitates the voice of a real baby harp seal." It only seems natural that people would develop attachments to such entities. And again, this is independent of their "actual" moral status.

Seeing as these reactive attitudes are a product of our natural sociality, an extension of this way of thinking is to consider how we feel when others have done something blameworthy or praiseworthy. In such cases, it seems that these individuals have gone beyond some expectation we have of them. Specifically, we judge their *quality of will* to have been supererogatory (or poor, depending on the kind of action). In other words, we have a belief that some or other expectation that we hold another to has been exceeded or breached (Wallace [32]: 11). This is essentially an evaluative *stance* that we adopt to others. To take such a stance is to believe that should some expectation be violated (even hypothetically) it would be *appropriate* for us to feel these emotions we call the reactive attitudes (resentment, guilt, etc.). Again, essential to Strawson's account is that our reactive attitudes are sensitive to others *quality of will* towards us, as this is realized outwardly in their behaviour ([22]: 7).

Strawson's theory is that our responsibility responses and our non-responsibility responses can be captured by appealing to the quality of an agents will. For example, imagine that Liana is on a flight and the plane encounters turbulence. In one scenario, the person sitting next to her spills a drink on her, due to the effects of the turbulence. In a second scenario there is no turbulence, but the same outcome occurs (Liana is covered in coffee). However, in this case, the person next to her makes eye contact, smirks, and then pours their drink on her. In the first case, resentment is inappropriate as the agent who spills the drink has no intention to cause harm, and so they have not violated any demand we may reasonably expect of them. In the second case, however, we are

[1] See http://www.parorobots.com/.

justified in our feelings of resentment, due to the poor quality of will they display. Their intention was clearly to mess coffee on a fellow passenger, and this renders resentment appropriate as a judgement of the poor quality of the agents' will. This, therefore, explains our responsibility responses to poor quality of will, but what about our *nonresponsibility* responses, cases where we *suspend* our reactive attitudes.

In the case above where turbulence causes the drink to spill, we agree that it is not problematic that there is nobody to hold responsible. Our nonresponsibility responses are justifiably redacted, given that the cause of the mess was not an intentional agent. However, in the case where the person intentionally messes their drink, we do indeed want to make an ascription of responsibility, and any successful theory of responsibility should be able to explain why this is so. Additionally, there is more than one way in which a redaction of our responsibility responses may be cashed out. In later section I will go into this in more detail, but for now it is enough to note that depending on whether we are dealing with abnormal circumstances (such as turbulence) or abnormal agents (such as small children), we may be justified in suspending our reactive attitudes ([22]: 8). Cases such as this are especially important considering their direct bearing on questions of responsibly-gaps. As I will argue, AI systems meet various *exemption conditions*.

3 Exemptions and excuses

In what ways might we find an agent to be an "unfitting" target of our reactive attitudes? There are two main ways we can think through this question: by considering *circumstances* or diverse *agential conditions*. Let us look at exempting circumstances first. Consider an example borrowed from Bernard Williams [34]. He asks us to imagine that a lorry driver "through no fault of his, runs over a child" ([34]: 28). Here, it is clear that the *cause* of the moral harm is the lorry driver, but we would not expect the parents of the child or anybody else to *blame* him or hold him morally responsible for the death of the child. We acknowledge that there are circumstances beyond our control that come to have an influence on our status as moral agents. In this example, the agent who caused the accident cannot be *blamed* for having done so. Importantly, however, even in this case we expect the driver to experience remorse or regret at his involvement in this tragic event. He might apologize to the family, try and console them, etc. We might also expect him to experience *agent-regret*, by holding *himself* responsible ([26]: 4). Thus, even though we do not find him responsible in the accountability sense, it seems reasonable that we expect him to feel remorse.

In the example above, we find that the driver is excused from responsibility. These are known as "type-1 pleas". In such a case, a moral harm has been committed, and although the lorry driver is the cause, type-1 pleas "tend to suspend our reactive attitudes for some particular injury in virtue of blocking any connection between their agents' behaviour and their qualities of will" ([22]: 8). The focus is on the particular circumstances in which the agent is embedded, as opposed to any fault in the agent. In these cases, we do not question whether the agent is a full moral agent or not, but rather we judge the circumstances in which the agent was embedded to be of such a nature as to provide a reason for a suspension of our reactive attitudes. However, there are cases where the capacity of the agent is indeed called into question, and these are known as type-2 pleas.

Type-2 pleas are concerned with situations in which we have unproblematic circumstances, but abnormal agents who are *exempted* from responsibility ([33]: 232). With type-1 pleas, agents are excused, but their having the capacity for moral agency is not in question. With type-2 pleas, however, it is the very nature of the agent that is called into question. In these cases, we acknowledge that some moral agents are exempted from our usual responsibility practices [22]. Examples include our treatment of psychopaths [16, 29], those with morally deprived upbringings, or children. In each of these cases (and there are many more), it is claimed that our usual responsibility practices are unfitting, given the constitution of the agent under review. When we exempt agents, however, what exactly are we doing? It is all very well to say that we suspend our usual responsibility responses, but what do these consist in?

3.1 The objective attitude

In exemption (or type-2 cases) we adopt what Strawson terms the "objective attitude", which commits us to treat agents "as creatures who cannot deserve praise or blame" ([23]: 323). Essentially, these agents become worthy of consideration in terms of *social cohesion*: we hold them "responsible" in ways that maximise social utility, as opposed to treating them as genuinely responsible agents. Take the example of small children: we punish them not because we think they are really "deserving" of punishment, but rather because we believe that such "punishment" will serve a useful purpose in their moral development. This punishment is independent of whether they "deserve" to be punished. Children, and small children especially, are in the process of developing, both morally and cognitively, and our holding them responsible is a pragmatic decision which aids in this process. In a sense, then, we regard them as natural objects, who are not fully morally responsible for their character or behaviour (yet!). Our "punishing" them comes not from considering whether they deserve it or not, but rather from considering the potential *consequences* of not doing so.

Spooling back the reel to type-1 and type-2 pleas, as these apply to AI, it seems as though type-2 pleas are the ones at issue. Specifically, when we find a machine to have performed some moral or immoral act and attempt to hold it responsible, we find that it is the wrong kind of agent to be a fitting subject of our responsibility ascriptions (at least for now).[2] I have already shown that this does not lead to a "gap" in responsibility, but establishing this does not tell us in which cases AI should be *exempt* from our responsibility practices wholesale. That is, while there might be no responsibility gap in a strict, philosophically narrow sense, this does not preclude such a gap emerging in a broader, social sense.

It might still be natural to feel angry at a self-driving car, or, at the very least, to feel angry that there is nobody to legitimately blame, and so now I will investigate the justifications we may have for suspending our reactive attitudes (and therefore adopting the objective attitude) in the face of agential AI systems. Due to the pluralism of responsibility, there may be *marginal cases*, where we show *ambivalence* in our treatment of agents. This occurs in our treatment of *natural* moral agents, and so I will first provide some detail as to how this "responsibility from the margin" plays out with respect to humans, and then apply it to AI [22].

4 The tripartite account of responsibility

According to the Strawsonian account above, our responsibility and non-responsibility responses are always a response to an agent's quality of will. However, as argued by David Shoemaker, having a "pure" interpretation of will fails to account for our *ambivalent* responsibility responses in many marginal cases. Consider the case of psychopaths. On the quality of will story, some might say that psychopaths are *excused* from our responsibility responses, due to them being incapable of having a proper quality of will (due to their cognitive or emotional impairments). But even if this is true, it still seems that we respond to the horrible behaviour of psychopaths with *disdain* or *contempt*, and these do seem to be responsibility-responses. So what is going on here? According to Shoemaker while we do not hold psychopaths responsible in the *accountability* sense (by, for example, realizing that resentment might be inappropriate given the empirical reality of psychopathy), we can still hold them responsible in the *attributability* and *answerability* senses of responsibility ([22]: 15).

Even though we might exempt psychopaths from accountability, we still believe them to be deficient in their *character* (attributability) and we expect them to be able to provide *reasons* or *justifications* (answerability) for the actions they have performed. To claim that *all* our responsibility practices flow from only an agent's quality of will, therefore, is too narrow. We are sometimes *ambivalent* in our responses, which means that we often have a combination of attitudes that might be in tension with one another towards the same agent, and this is especially true in the case of psychopaths (and those with poor formative circumstances). It does seem that some of our responsibility practices are justified, while others are not. Thus our responsibility responses can take various, distinct forms, to unique agential conditions. This is especially illuminating in the case of AI systems, as a key question is whether such systems are agential "enough" to exclude them from our responsibility practices. That is, is their responsibility *mitigated* or are they *exempt*? In the next section, I will explore these two questions and see what light it might shed on our relationship to AI.

5 Exempting and/or including AI

5.1 Attributability

Responsibility as attributability is usually focused on our responses to the faults or excellence of others ([22]: 38). The standard pairing of reactive attitudes here is *admiration* and *disdain*. However, while I might "admire" Table Mountain in Cape Town, this does not seem to track the kind of admiration that is going on in our responsibility responses. Admiration, in the responsibility sense, is concerned with the quality of the *character* of the *agent* in question. Our admiration of an agent seems to consist in a positive evaluation of who they are and what they stand for. Conversely, we might feel *disdain* towards an agent. Here we would be tracing their actions and attitudes to some fault or vice in their character. Importantly, in both admiration and disdain, there are conditions under which it would be inappropriate to attribute certain aretaic responses to an agent. For example, if your partner is usually very caring and considerate, but has recently come under a lot of stress at work leading them to snap and raise their voice at you, you would not disdain them. Their outburst would be *out of character*, and thus would not be reflective of the values that they *really* stand for ([22]: 42). These values, as hinted at above, would be linked to the agents' cares and commitments, as expressed in their attitudes. For an agent to be attributability-responsible for an attitude of theirs, therefore, it must be the case that this attitude *expresses* some part of the agents' cares, commitments, or "commitment clusters" ([22]: 59). As I argued previously, however, machines and currently existing AI,

[2] This would be the case for both our blaming and praising practices.

however, cannot be said to "care" about anything. They do not have character in the sense required, as the values they "stand for" (if any) are a function of those natural moral agents who designed and deployed them. While we often make such aretaic judgements about artifacts ("why won't my car start! It's clearly doing this on purpose"), we know that such attributions are incoherent (and therefore not fitting), and the same is true for such sentiments towards AI.

5.2 Answerability

Responsibility as answerability is concerned with quality of *judgement* ([22]: 65). Here the associated reactive attitudes would be regret or pride. Normally, these sentiments are experienced by the agent over some decision they may have performed. In the case of regret, the typical thought would be something like "if only I had done otherwise", with a certain action tendency towards changing how one came to the faulty decision in the first place and improving one's decision-making in the future. In the case of pride, we would have the inverse, with the associated thought something like "I did the right thing", with an action tendency towards maintaining the kind reasoning process that led to the good decision ([22]: 66). How exactly might AI be exempted from our responsibility responses in this case? The reason for this exemption is that AI based systems cannot engage in the *right kind of reasoning*.

One way to deal with this problem can be found in the push towards *explainable* AI (Van de Poel [30]). Explainable AI here can refer to a number of things, but it essentially means that, when designing AI systems, they ought to be created in such a way that should we wish to investigate the system, we would be able to trace the causal mechanism by which it came to decide on a particular course of action. This is no easy task, especially for deep learning algorithms "which are developed with the goal of improving functional performance. This results in algorithms that fine-tune outputs to the specific inputs, without giving any insights on the structure of the function being approximated" (Dignum [9]: 88).

However, certain AI systems (for example, medical diagnostic tools) come to feature in the judgements made by human beings. Doctors might recommend a particular course of action based on the "advice" given to them by an AI. Does this "advice" constitute a kind of judgement? Not at all. Due to AI not having the ability to do anything more than provide technical explanations for its behaviour, *they do not have judgment at all*, and so asking questions about the *quality* of their judgement is a non-starter.

While of course these technical explanations can come to feature in the judgements of human agents who evaluate various algorithmic systems, there is no gap in responsibility, and to think there is would be kind of category mistake. When we talk about explanations (with respect to algorithmic systems) what we are interested in is a transfer of knowledge from the system to some human agent (usually, an expert in the relevant field), who can then make proper use of the information. Explanations depend on the system, and merely describe what is going on. Thus we can see such explanations as being, in a sense, *intrinsic* (Henin and Le Métayer [13]). Justifications, on the other hand, are *extrinsic*, in that our evaluation of them is dependent on whether they are appropriate or not, and to do so we make appeals the concepts and theories that are *outside* of the algorithmic system itself (Henin and Le Métayer 2021). For example, with respect to the rule of law, we not only require explanations but also justifications that are deemed legitimate within that particular context. This is not to say that all justifications even *require* an explanation. It is possible to have a justification that makes no mentioned of the underlying logic of a specific AI system, just as explanation by itself does not imply justification. These systems are thus exempt from our answerability responsibility practices, even if they come to inform how such practices might apply to natural moral agents (in the form of AI systems being sources of various forms of information).

However, due to the ubiquity of AI systems (as noted earlier), we find ourselves consistently interacting with them. And so the line of thought I want to investigate is whether we might not demand *justifications* from them, but perhaps more modestly, simply "answers". Now, as I have detailed, AI cannot form judgements and therefore provide justifications for their actions. Nevertheless, might there be a sense of what Daniel Tigard calls "technological answerability" that machines could achieve (2021)? Technological answerability is "a capacity in technological systems for recognizing human demands for answers and responding accordingly" ([27]:10). What Tigard aptly suggests is that "given the increasing ubiquity of sophisticated technologies in our daily lives and the fact that we might not be able to discern reasons for a system's behaviour, efforts to increase technology's answerability might solve some problems, even if it creates others" (2021: 5). Such "answers" could be very simple. For example, you might be coming to the end of a particular show on Netflix, and as such the platform helpfully suggests other shows you might be interested in. There might be many causes as to why you received this list of shows, including but not limited to your past viewing history, corporate sponsorship deals Netflix might have, the probability that you will binge the shows, etc. Now, a call for technological answerability would simply demand that the system, should you as a user query it, be able to provide this information. This need not be some detailed technical explanation at the level of how the algorithmic system works, but rather some information that helps the user understand the immediate cause of the

system's output ([27]: 13). Such "answers", therefore, need not be satisfactory from the perspective of moral responsibility but rather from the perspective of human psychology. Such a design requirement might also reduce the *perceived* threat of responsibility gaps if people are indeed satisfied by these answers. Of course, on the flip side, this also raises worries about deceit: it is entirely plausible that companies would provide answers satisfying to human psychology to gloss over important ethical issues in the design of their systems. However, my point here is not to argue in favour of such answerable machines, merely to posit the possibility, and how this might influence our responsibility practices. Next I turn to responsibility as accountability.

5.3 Accountability

The associated reactive attitudes here are usually resentment or indignation, as this is the way Strawson famously framed his argument. Resentment is what I feel towards you for having slighted me, whereas indignation is what I feel on behalf of another who I believe has been slighted. These two can therefore be understood together, as they essentially involve a reaction to a perceived instance of poor treatment. In my discussion, therefore, I will instead make use of the language of "agential anger" to capture both resentment and indignation. Framing the discussion in terms of agential anger also has some conceptual upsides.

When my phone goes on the fritz I often get angry at it. Why? Well, perhaps my anger is due to the frustration of some goal I had in mind when using my phone. Maybe I wanted to buy the latest running shoes (helpfully recommended to me by my choice of digital oligarch on their mobile application) but just as I wanted to place my order, my phone reported an error, and I subsequently missed out on the special. I had a goal, and my phone frustrated my ability to achieve that goal. My anger, on this account, might be *justified* (in the sense that I really did want those shoes and my phone was the reason I now cannot get them) but would my anger also be *fitting* or *legitimate*? If the object of anger is goal frustration, then it seems my anger is fitting. However, as noted above, we are concerned here with *agential* anger. Agential anger is not merely about goal frustration but has an action tendency that motivates those experiencing it to take *revenge* or *retribution* against an *agent*. Thus, agential anger towards my phone is absurd because it would make no sense to take revenge on my phone for what happened. While it might be cathartic to throw the phone against the wall or on the floor, the phone is not an agent in the appropriate sense, and is thus not a proper target for such responses. In the case of AI, then, it is not enough that we merely feel angry at these systems, our anger must also be fitting. If it is not fitting, then we need a justification for why we ought to suspend our reactive attitudes.

The question now becomes what exactly agential anger is responding to. That is, what might render it reasonable to suspend our agential anger? The most plausible story here is that when I get angry at another agent "I am lodging some sort of complaint or demand", where I consider the agent in question to have failed to consider the potential consequences of their intentional action ([22]: 93). Fitting anger is therefore a response to poor *quality of regard* in an agent ([22]: 112). Regard here refers to the ability of an agent to weigh and take into account the *interests* of others. As currently existing AI systems lack the ability to weigh interests in this way, they cannot display poor quality of regard and are therefore exempted from responsibility as accountability.

However, responsibility as accountability is unique out of these three senses of responsibility in that it has a built-in confrontational element ([22]: 87). That is, the agent who has been slighted should be able to *explain* why they feel hard done by, and the agent who is being held to account should be able to *understand* and appreciate that they have done something wrong. This feature of accountability can be seen in the example of psychopathy outlined earlier: due to psychopaths lacking the ability to appreciate *why* what they have done is wrong, they are exempted from responsibility as accountability. In other words, their inability to take up the normative perspective of those they have harmed, due to the type of agents that they are, exempts them. Similarly, advanced AI systems are also incapable of taking any form of "perspective" and are also exempted. Naturally, the question of whether AIs could be phenomenally conscious rears its head. However, for my purposes, it is enough that none are conscious *yet*. The further metaphysical question of whether or not they *could* be is not one I aim to settle here, nor does it have much bearing on my argument.

6 Conclusion

My aim in this paper was to address the topic of responsibility gaps form a broader perspective. I have shown that while AI-based systems certainly threaten our responsibility practices, they do not undermine them. Specifically, I argued that in each sense of responsibility, we can appreciate the risks posed by AI systems and attempt to safeguard against them. By bringing standard work in moral philosophy, such as that of Strawson, to bear on questions related to AI, I hope to have shown that there is still much that can be gained from traditional philosophical concepts, especially as these are applied to novel and emerging technologies. The upside of the argument I have presented is that it does not necessarily turn on any demanding metaphysical questions, and instead offers a philosophically robust *pragmatic* approach to demanding normative questions in the ethics of AI.

Funding Open Access funding enabled and organized by Projekt DEAL. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project 254954344/GRK2073 "Integrating Ethics and Epistemology of Scientific Research".

References

1. Bernáth, L.: Can autonomous agents without phenomenal consciousness be morally responsible? Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00462-7
2. Buolamwini, J., Gebru, T. : Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of Mahcine Learning Research. vol 81, pp. 1–15 (2018). https://doi.org/10.2147/OTT.S126905
3. Champagne, M., Tonkens, R.: Bridging the responsibility gap in automated warfare. Philos. Technol. **28**(1), 125–137 (2015). https://doi.org/10.1007/s13347-013-0138-3
4. Coeckelbergh, M.: Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI & Soc. **24**(2), 181–189 (2009). https://doi.org/10.1007/s00146-009-0208-3
5. Coeckelbergh, M.: Moral appearances: emotions, robots, and human morality. Ethics Inf. Technol. **12**(3), 235–241 (2010). https://doi.org/10.1007/s10676-010-9221-y
6. D'Arms, J., Jacobson, D.: Sentiment and value. *Ethics*, 110(4): 722–748 (2000). Available at: https://www.jstor.org/stable/10.1086/233371%0AJSTOR.
7. D'Arms, J., Jacobson, D.: Anthropocentric constraints on human value. In: Shafer-Landau, R. (ed.) Oxford studies in metaethics, vol. 1, pp. 99–126. Oxford University Press, Oxford (2006). https://doi.org/10.1093/oso/9780198859512.001.0001
8. Danaher, J.: Robots, law and the retribution gap. Ethics Inf. Technol. **18**(4), 299–309 (2016). https://doi.org/10.1007/s10676-016-9403-3
9. Dignum, V.: Responsible artificial intelligence. Springer Nature Switzerland, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6
10. Frankish, K., Ramsey, M.W.: Introduction. In: Frankish, K., Ramsey, M.W. (eds.) The Cambridge handbook of artificial intelligence, pp. 1–11. Cambridge University Press, Cambridge (2014)
11. Gebru, T.: 'Race and Gender', In: Dubber, M., Pasquale, F., and Das, S. (eds.) Oxford Handbook of the Ethics of AI. New York: Oxford University Press, pp. 253–270 (2020)
12. Gunkel, D.J.: Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf. Technol. (2017). https://doi.org/10.1007/s10676-017-9428-2
13. Henin, C., Le Métayer, D.: Beyond explainability: justifiability and contestability of algorithmic decision systems. Ai Society. (2021). https://doi.org/10.1007/s00146-021-01251-8
14. Lauwaert, L.: Artificial intelligence and responsibility. AI & Soc. (2021). https://doi.org/10.1007/s00146-020-01119-3
15. List, C.: Group agency and artificial intelligence. Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00454-7
16. Litton, P.: Responsibility status of the psychopath: on moral reasoning and rational self-governance. Rutgers Law J. **39**(349), 350–392 (2008)
17. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. **6**(3), 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
18. Nyholm, S., Gordan, J.-S.: Ethics of artificial intelligence. In: Fieser, J., Dowden, B. (eds) Internet Encyclopedia of Philosophy (2021). https://doi.org/10.5860/choice.191051
19. Orr, W., Davis, J.: Attributions of ethical responsibility by artificial Intelligence practitioners. Inf. Commun. Soc. **23**(5), 719–735 (2020). https://doi.org/10.1080/1369118X.2020.1713842
20. Russell, S., Norvig, P.: Artificial intelligence. In: Russell, S., Norvig, P. (eds.) A modern approach, 3rd edn. Prentice Hall, Boston (2010)
21. Ryland, H.: Could you hate a robot? And does it matter if you could? AI & Soc. (2021). https://doi.org/10.1007/s00146-021-01173-5
22. Shoemaker, D.: Responsibility from the Margin. Oxford University Press, Oxford, United Kingdom (2015). https://doi.org/10.1016/j.cirp.2016.06.001%0
23. Sommers, T.: The objective attitude. Philos. Quarterly (2007). https://doi.org/10.1111/j.1467-9213.2007.487.x
24. Sparrow, R.: Killer robots. J. Appl. Philos. **24**(1), 62–78 (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x
25. Strawson, P.: Freedom and resentment. Proc. British Acad. **48**, 1–25 (1962)
26. Tigard, D.W.: There Is no techno-responsibility gap. Philos. Technol. (2020). https://doi.org/10.1007/s13347-020-00414-7
27. Tigard, D.W.: Technological answerability and the severance problem: staying connected by demanding answers. Sci.Eng. Ethics (2021). https://doi.org/10.1007/s11948-021-00334-5
28. Tollon, F.: Designed to seduce: epistemically retrograde ideation and YouTube's recommender system. Int. J. Technoethics **12**(2), 60–71 (2021). https://doi.org/10.4018/IJT.2021070105
29. Tollon, F.: Do others mind? Moral agents without mental states. South African J. Philos. **40**(2), 182–194 (2021). https://doi.org/10.1080/02580136.2021.1925841
30. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Mind. Mach. **30**(3), 385–409 (2020). https://doi.org/10.1007/s11023-020-09537-4
31. Verdiesen, I., Santoni de Sio, F., Dignum, V.: Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. Mind. Mach. **31**(1), 137–163 (2021). https://doi.org/10.1007/s11023-020-09532-9
32. Wallace, R.J.: Responsibility and the moral sentiments. Harvard University Press, Cambridge, Massachusetts (1998). https://doi.org/10.2307/2956371
33. Watson, G.: Responsibility and the limits of evil: variations on a Strawsonian Theme. In: Shoeman, F. (ed.) Responsibility, character and the emotions: new essays in moral psychology, pp. 256–286. Cambridge University Press, Cambridge (1987)
34. Williams, B.: Moral Luck: philosophical papers 1973–1980. Cambridge University Press, London, England (1981). https://doi.org/10.5840/intstudphil198517175