Philosophy & Technology (2021) 34:589–607
https://doi.org/10.1007/s13347-020-00414-7

RESEARCH ARTICLE

There Is No Techno-Responsibility Gap

Daniel W. Tigard[1]

Received: 22 January 2020 / Accepted: 23 June 2020/ Published online: 9 July 2020

Abstract
In a landmark essay, Andreas Matthias claimed that current developments in autonomous, artificially intelligent (AI) systems are creating a so-called responsibility gap, which is allegedly ever-widening and stands to undermine both the moral and legal frameworks of our society. But how severe is the threat posed by emerging technologies? In fact, a great number of authors have indicated that the fear is thoroughly instilled. The most pessimistic are calling for a drastic scaling-back or complete moratorium on AI systems, while the optimists aim to show that the gap can be bridged nonetheless. Contrary to both camps, I argue against the prevailing assumption that there is a technology-based responsibility gap. I show how moral responsibility is a dynamic and flexible process, one that can effectively encompass emerging technological entities.

Keywords Machine ethics · Artificial intelligence · Moral responsibility · Moral agency · Blame · Human-robot interaction

1 Introduction

In what can now be regarded as a landmark essay in the ethics of technology, Andreas Matthias proclaimed that current developments in autonomous, artificially intelligent (AI) systems are creating a so-called responsibility gap. With the emergence of machines that are capable of operating by way of unfixed rules and learning new patterns of behavior, it is said, our ability to trace responsibility back to the manufacturer or operator has come under threat. Indeed, the gap is allegedly "ever-widening" and should be understood as no small concern, as the problem stands to infect both "the moral framework of society and the foundation of the liability concept in law" (Matthias 2004). And while any serious threat to our moral well-being or to fundamental notions in our legal systems should be quite concerning, we must ask ourselves:

✉ Daniel W. Tigard
daniel.tigard@tum.de

[1] Institute for History and Ethics of Medicine, Technical University of Munich, Munich, Germany

How severe is the threat posed by emerging technologies? Should we be afraid that machines will utterly constrain our ability to locate responsibility?

A great number of authors, responding more or less directly to Matthias's responsibility gap, have indicated that the fear is thoroughly instilled. We should be quite afraid, many think, that technology is posing a unique and potentially insurmountable obstacle to our usual sense of morality and law. The questions, then, to be addressed are: What should we do about the ever-widening responsibility gap? What, if anything, *can* be done about it? It is here that we see the most common division in the recent literature. According to some authors, the gap simply cannot be bridged. The use of AI, for some, will leave us without anyone—or anything—to plausibly hold to account for harms and damages caused by such technologies. Given the importance of locating responsibility, the thought goes, we must drastically scale back or altogether cease our deployment of AI systems. For the sake of brevity, this camp can be referred to, roughly, as the techno-pessimists.[1] Granted, it is easy to share their concerns, particularly when considering the use of autonomous technology in such high-stake domains as medical practice, warfare, and wider political decision-making (e.g., Sharkey 2010; Asaro 2012; Char et al. 2018; Danaher 2016a).

On the other side, however, there are those who argue that the gap can be bridged. The techno-optimists, as I will call them here, are those who would prefer to harness the newfound benefits of technology and proceed with its deployment. With this overarching agenda, the burden they accept is to show how, in the face of the potentially widening gap, we can locate responsibility nonetheless. Where the optimists differ among themselves is typically on the various proposals for how responsibility can or should be found (e.g., Marino and Tamburrini 2006; Hanson 2009; Rahwan 2018; Nyholm 2018).

Importantly, the techno-optimists and the techno-pessimists appear to *agree* upon a fundamental premise, namely that technology poses an especially unique problem for our existing moral and legal practices. For this reason, I propose we step back and ask whether or not there is a responsibility gap in the first place. I also want to focus the present discussion primarily on moral responsibility, rather than on legal notions (such as liability). No doubt, there are important lessons to be drawn in both directions, but I take it that the gap in moral responsibility is the more puzzling inquiry. After all, regulatory gaps can in principle be repaired with new regulation—whether hard law concerning data protection or corporate self-regulation[2]—while potential gaps in moral responsibility are far less clear, and thereby less easily resolved. In order to clarify the domain-specific nature of the issue, I will refer to the unique threat to moral responsibility posed by emerging technologies as the techno-responsibility gap, for surely we encounter difficulties in identifying responsibility when it comes to non-technological harms as well.[3] In this essay, I aim to challenge the prevailing assumption maintained by both the techno-optimists and the techno-pessimists. As my title reveals, I will argue that there is no techno-responsibility gap.

[1] Note, I use "techno-pessimism" (and "techno-optimism") rather loosely, as such labels are applied with various connotations in a range of contexts. Indeed, it appears that a techno-pessimist might be one who is certain of an impending robot apocalypse or who regularly expresses gripes over the latest smartphone. My usage will remain general but leans toward the former.

[2] See, for example, Wachter and Mittelstadt (2019), Dignam (2020), and Morley (manuscript).

[3] Consider genuine moral dilemmas or cases of moral luck (e.g., Williams 1981). I address such cases in more detail below.

To begin, I must first make clear exactly what is meant when positing the existence (or non-existence) of responsibility gaps. Here, also, I will expand upon the various reasons to be pessimistic or optimistic about the prospects of overcoming such gaps when they supposedly result from technology. Next, I shift from the question of whether or not techno-responsibility gaps can be bridged to whether or not they exist, an inquiry which should be theoretically prior, as I see it. I then invoke pluralistic conceptions of moral responsibility, in order to expand upon a recent work arguing against gaps in accountability (Köhler et al. 2017). With this broader account in hand, I argue that responsibility, as a dynamic and flexible process, can effectively encompass emerging technological entities.

2 Techno-Responsibility Gaps and Our Bridging Prospects

In order to clarify speculation about the existence of techno-responsibility gaps, a fruitful basis of investigation will be to make clear what it means to face a gap in responsibility, that is, without the technology-based qualification. Once these notions have been established, I will explain in greater detail why we might be especially pessimistic or optimistic about overcoming the often alarming technological variety.

2.1 Gaps in Responsibility: from Mundane to Extraordinary

Starting with the most elementary observations, it can safely be said that our lives are rife with noticeable events, some of which we enjoy and find beneficial, while others are disagreeable and harmful. In either case, it serves us well to understand the sources of the various effects upon us, as we naturally aim to promote the positive and avoid the negative. Our responses, as it were, often take the form of praise or blame, instances of which will of course vary greatly in their degrees, their functions, and their inner and outward manifestations. Ever since P.F. Strawson's renowned essay encouraged ethicists to largely put aside the concerns over free will and the threat of determinism, numerous authors have come to see moral responsibility as a social function of the so-called reactive attitudes. These are, in Strawson's words, the "natural human reactions to the good or ill will or indifference of others towards us" (Strawson 1962).

But what happens when the sources of harms or benefits cannot be clearly identified, either because there are no identifiable sources or simply because we are ignorant of them? That is, what happens when there is no obvious target of our praise or blaming attitudes? Consider, for example, a relatively mundane case.[4] John walks the same stretch of sidewalk every day to get to work, and most often the sidewalk is pleasantly neat and free of litter. On a particularly grueling Monday, he walks his usual path, which had recently been swept. Even so, John steps in freshly chewed gum, mucking up his favorite shoes. Naturally, John is angered. He then sees the city sweeper up ahead and quickens his pace to go have a word. The sweeper informs that the block was recently swept and that these things happen. John realizes his anger is misplaced, so he looks around for the gum-chewer. However, no one else is in sight.

[4] For helping to inspire the following example, I am indebted to Nate Stout and his paper "Blame *De Re* and *De Dicto*" (manuscript).

Plausibly, given the prevalence of mundane cases such as John's gum-shoe, any one of us might face what we can think of as a *gap* in responsibility. That is, the subject of some effect recognizes the event, perceives it to some degree as a harm or benefit, and responds naturally in a way that would ascribe a sort of morally significant credit to the source.[5] However, the source is absent, unknown, or perhaps entirely nonexistent, leaving the subject's reaction without an appropriate target. Alternatively, where the source is known and present in a way that allows the subject's responses to find their appropriate target, it appears there is no gap in responsibility.

Importantly, where the source is a fully functional adult moral agent, the subject's assignment of moral credit can be said to be "fitting" (D'Arms and Jacobson 2000). In other words, the natural reactions maintained by the subject are correctly identifying the source as such. For instance, Jane's resentment at a friend whom she believes intentionally slighted her is fitting where her friend in fact intentionally slighted her. By contrast, where the source is present but is not a fully functional adult moral agent, the assignment of moral credit is simply inappropriate or "unfitting," as the subject's response incorrectly identifies some individual as the source. This sort of inappropriateness is displayed, for example, when we cringe at the sight of a parent excessively berating a small child. Along with psychopaths, persons with intellectual disabilities or a morally deprived upbringing (among others), children are often thought to be exempted from our usual, full-fledged responsibility practices (Shoemaker 2015).

Does this mean that where the immediate source of harm *is present*, but cannot be appropriately targeted with negative responses, we face a gap in responsibility just as we would where the source is absent? The all-too-common refrain to be stated here is: it depends—namely on what the source of the harm and any victims to it are prepared to do in response. Consider a more extraordinary case. Many readers will be familiar with Bernard Williams's famous example of moral luck. Briefly, in a tragic accident, a lorry driver "through no fault of his, runs over a child" (Williams 1981, p. 28). As Williams and others maintain, cases like this give us reason to accept that there are features of our lived experiences beyond our awareness or control, which inevitably affect our status as responsible agents. But without belaboring the debate over moral luck, I want to look briefly at what these complexities can tell us about the apparent presence of a gap in responsibility.

Imagine, on one hand, that the lorry driver responds to the accident with something like regret; he apologizes profusely to the parents of the child, and so on. Indeed, this scenario resembles what Williams had in mind with his introduction of "agent-regret." Here, what is important to notice is that the driver is holding *himself* responsible.[6] In other words, there is no gap in responsibility, since the driver experiences the harmful event in a way that allows his responses to find their appropriate target, namely himself. And granted, because he is not morally at fault for the harm, there is a sense in which he is not worthy of blame—after all, while the parents and any bystanders might be upset, they cannot reasonably blame him. Still, the driver's natural, self-directed blame is nonetheless appropriate, for imagine the very same scenario but where he recognizes his innocence and shamelessly shrugs off the event.

[5] By "morally significant credit," I do not mean to defend desert-based accounts of responsibility. Those with other, perhaps, consequentialist leanings are invited to substitute their preferred account.

[6] Of course, this claim and, generally, the nature of agent-regret have been subject to volumes of debate. I strongly recommend the analysis offered by Jacobson (2013).

On the other hand, then, consider the tragic accident where the lorry driver in fact experiences nothing akin to regret or remorse (or agent-regret). Surely, here is where the parents or any bystanders might reasonably blame the driver. But their blaming responses are appropriate only insofar as they are directed at the driver's shamelessness or insensitivity. In other words, he can be appropriately blamed by others for not blaming himself, but not for the event itself, despite the fact that he is—in a nontrivial sense—the source of the harm. Thus, given that there is no appropriate target for the blaming responses of those who experience such reactions, it seems that there is a gap in responsibility. In short, the driver *alone* is in a position to hold himself responsible, yet he fails to do so.[7]

With this scenario, we see a crucial addition to the notion I offered above. That is, there are gaps in responsibility where one recognizes some harm (or benefit) and responds naturally in a way that would ascribe a sort of moral credit to the source; however, this response is without an appropriate target, given that the source is absent, unknown, nonexistent, or is not a fitting target. Notice here that the lack of fittingness might result from the source being either *excused* from responsibility or *exempted* from it. Cases of the former sort are known as "type-1 pleas"—where the agent is normal, but the circumstances are abnormal in ways that inhibit our reactive attitudes.[8] Consider again that the lorry driver is a fully functional adult moral agent but, due to extraordinary circumstances, played a causal role in a tragic accident. By contrast, cases where one is exempted from responsibility—"type-2 pleas"—are where the circumstances are normal, but the agent is abnormal in ways that inhibit our reactive attitudes. Here, again, consider cases of apparent wrongdoing involving children, psychopaths, or persons with intellectual disabilities. Considering the excuse-exemption distinction, it appears that type-2 abnormalities are the kinds of cases Matthias and others typically have in mind when stoking our fears of an ever-widening techno-responsibility gap. I will expand upon this suggestion.

2.2 Techno-Responsibility Gaps

If indeed we see gaps in responsibility resulting from absent, unknown, or nonexistent sources of harm, as well as from sources that are not fitting targets of our reactive attitudes, it may be that responsibility gaps are commonly encountered. Do we, then, face gaps in responsibility as a result of increasingly sophisticated technologies? On the account I have offered so far, to say that gaps can result from AI is simply to maintain that one is responding naturally to some harmful event in way that would ascribe blame to the source; however, the source, while perhaps present and known, is not an appropriate target of one's responses.[9] What exactly is it that distinguishes cases of AI from non-technological gaps in responsibility?

Picture the not-too-distant future where a self-driving lorry, through no fault of its own, runs over a child. While some projections foresee self-driving lorries primarily on highways, we can imagine cases where careless children cross their paths nonetheless,

[7] I discuss this phenomenon, and its relation to moral distress, in Tigard (2019a).

[8] For a fuller explanation, see Watson (2004, pp. 227–228) and Shoemaker (2015, pp. 7–8).

[9] One reason for thinking this is that AI is not a moral agent. Still, despite their usual concurrence, I believe questions of *responsibility* in technology can be addressed apart from questions of *agency*. I discuss this further in Tigard (2020).

in a way that the fault cannot be ascribed to the lorry or its operating company. Imagine that before it is known that the lorry was without a human driver, the child's family responds naturally with anger. Their immediate reactions would ascribe moral responsibility to the source; however, that source is not a fitting target. At first glance, the case appears to bear close resemblance to Williams's example. The key difference, it seems, is that in the classic example, we see the possibility of the driver holding himself morally responsible. Unlike even the most intelligent machines, a human could truly take responsibility—by responding with guilt or regret—in order to help the family find solace, begin moving forward, and so on (Mason 2019; Tigard 2019b). By contrast, in the self-driving case, the lorry "driver" is incapable of exercising self-directed responses like guilt or regret. In this way, it seems that although machines can certainly cause harm, they cannot be held responsible, either by themselves or others, for they cannot display good or ill will.[10] As a result, the family would likely modify their responses upon learning of the self-driving lorry. These observations might seem rather obvious, yet they may support the idea that where technology causes our responses to be without a fitting target, such as the family's initial anger at the self-driving lorry, we face a gap in responsibility.

Admittedly, here it appears that the techno-responsibility gap is indeed a theoretical likelihood and a great practical concern. Still, in cases of purely technology-based harms—where we cannot trace responsibility back to the manufacturer or operator—we may nonetheless respond initially with our usual attitudes and practices. The significance of this fact is less a matter of our responses themselves. Instead, what we must take careful note of are the reasons for any modification in our initial attitudes. As I suggested concerning the classic example, where the child's family realizes that it was truly an accident of the lorry driver, their initial anger will be modified on type-1 grounds. Once the facts are known, that is, the driver will be excused from responsibility, as the family comes to see that he did not manifest any sort of ill will. He is, however, still the kind of being that is capable of possessing good or ill will.

In the case of the self-driving lorry, particularly in the near-future where such tragedies are still unusual, the family plausibly responds with something like the anger we direct at fellow moral agents. "Agential anger," as David Shoemaker (2015) dubs it, is contrasted with the sort of anger we typically display toward mere objects or perhaps fate—picture, for example, our anger at the office copy machine. Importantly, agential anger and related responses involve not just the feeling of a goal being frustrated, but also the thought that someone has culpably offended and a tendency to act accordingly, say, by communicating the offense or seeking retribution.[11] Why, then, are these features significant in response to self-driving lorries? Would we not simply find the family to be mistaken where they exhibit agential anger toward an AI system?

I suggested above that where the child's family comes to learn that the lorry driver is not human, their initial anger will be modified, but the modification is grounded in a

[10] With no good or ill will, it appears that machines are not suitable candidates for our responsibility practices, since the Strawsonian view relies on discerning one's "quality of will," as it is often put. While this feature alone causes difficulties for applying the quality-of-will approach to machines, later developments—as I discuss below—show responsibility to be a pluralistic enterprise. I thank an anonymous reviewer for pressing me on this point.

[11] This multidimensional view finds substantial support in moral psychology literature. See Oakley (1992), D'Arms and Jacobson (2006), Shoemaker (2015), and Szigeti (2015).

type-2 abnormality. That is, once the facts are known, the "driver" will be exempted from responsibility, given the family's realization that there could not have been any sort of ill will be displayed in the harm that came to their child. Like other sources which cannot appropriately receive the full force of our reactive attitudes, AI systems are simply not capable of possessing ill will toward us.[12] The point to be emphasized here is that while humans can be excused from responsibility when playing a causal role in tragic accidents, the main reasons we modify our attitudes toward AI systems playing causal roles do not concern the abnormal circumstances. Instead, we make adjustments to distinct agential conditions, which are immensely diverse. I will return to this notion but first expand upon why we might be optimistic or pessimistic about overcoming responsibility gaps supposedly resulting from technology.

2.3 Our Bridging Prospects

As I prefaced, some have found the techno-responsibility gap to be problematic but surmountable—*bridgeable* is, of course, an expected analogy. Differences between such accounts are seen simply in the materials. Clarifying whether or not the gap is bridgeable by any means should help to see why we might be optimistic or pessimistic about problems for responsibility resulting from technology.

In an effort to articulate "responsibility ascription policies" for science and technology, Dante Marino and Guglielmo Tamburrini (2006) suggest that we can bridge the gap with individual computer scientists and engineers, along with their organizations. These actors can evaluate relevant risks and benefits, say, from machine-learning robots that might cause harm. They can help to identify in advance the damages that are deemed socially sustainable and the criteria for appropriately distributing liability for damages, even where "there is no clear causal chain connecting them to the damaging events" (Marino and Tamburrini 2006, p. 49). By developing clear rules and criteria, on this approach, the gap is bridgeable both retrospectively and prospectively with the help of scientists and engineers. Similarly, individuals who might be plausibly held responsible are those in *command* of a machine's behavior—picture military commanders or soldiers giving orders to military robots (Hellström 2013).

More recently, Iyad Rahwan (2018) proposes broad collections of individuals as the bridging materials, showing that society at large—and not just single humans—can be "in-the-loop" when programming autonomous technologies. Thomas Hellström (2013, p. 105) also notes that "society may decide to collectively share responsibility" for the behavior of machines such as military robots. Particularly in democratic societies, where the public may seem at least complicit in political decision-making (Archard 2013), we see a path to holding very large groups responsible for the development or use of certain technologies. On these accounts, even where machines learn and act upon rules not initially programmed, by keeping societal values in-the-loop, we keep society as a whole "on the hook" when things go wrong.[13]

[12] And I believe we need not seriously entertain the possibility, at least not yet. There are some who disagree, notably Bostrom (2014).

[13] Similarly, Taddeo and Floridi (2018) suggest *distributing* responsibility among the many actors that together produce effects via AI.

Still, it is important to notice that seeking a broader locus of responsibility often works against the project of bridging the gap. The problem of "many hands" as it is often called (e.g., van de Poel et al. 2012) is where a negative effect is produced collectively by multiple actors, none of whom can be assigned responsibility, since no single individual knew enough or intended or even had the power to cause the outcome in question. In a recent work, Roos de Jong (2020) raises such concerns for Sven Nyholm's account of human-robot collaborations. For Nyholm (2018, 2020), we must address pertinent questions, such as: Who is supervising? Who is currently in control? And whose preferences are guiding autonomous systems? As de Jong argues, these questions may well point us in different directions, since a single AI or robotic system can "simultaneously participate in more than one human-robot collaboration" (2020, p. 731). Nonetheless, as Nyholm can maintain, a human collaborator could and often should take responsibility for a system's behavior and, in this way, no matter how many hands produced the outcome, we do not lose sight of responsibility altogether.

In any case, whether bridges over the gap are built from individuals or collections, it appears on these accounts that something other than the direct source of harm is being held responsible. That is, individuals or collections can pay for damages their machines cause or promise to improve technology for the future. But notice that the victims to any harms, while likely better off than where they are left to suffer alone, are being compensated or reassured by someone only indirectly involved. As such, victims' initial reactions and attempts to locate responsibility in the immediate source will be left unsatisfied (Danaher 2016b).

Perhaps more satisfying, then, are the proposals of those who build their bridges out of human-machine composites. F. Allan Hanson (2009), for example, suggests that we hold "extended agencies" responsible. The idea here is that being a moral agent is simply not limited to individual humans. Just as a single human can act in conjunction with other humans, our actions are often enabled or disabled by inanimate objects, namely tools and technologies.[14] As such, it is said, we can assign responsibility *jointly*, that is, to humans and non-human components as a whole. However, it is one thing to accept that technology enables certain actions, which are then performed by humans who are *ultimately*, *morally* responsible. It is quite another to suppose that technological devices themselves can be morally responsible, even in part or as an accomplice, we might say. What exactly does it mean to say that tools (like firearms) are morally "complicit" in actions (like murders)? Additional difficulties of the composite view include explaining how to locate responsibility when the apparent source is a purely technological entity, a machine that learns and behaves entirely on its own. These inevitabilities are precisely why some have taken a more pessimistic outlook on bridging the techno-responsibility gap.

Consider, briefly, Robert Sparrow's (2007) widely discussed account of "killer robots." Under the principles of *jus in bello*, the ability to locate responsibility is required for engagement in a just war. The problem with autonomous technologies—AI weapon systems, among other applications—is that when things go wrong, "it is not

[14] Support for this idea is found in theories that ascribe a morally or politically significant collaborative role to technological artifacts (e.g., Winner 1980; Verbeek 2008; Nyholm 2018, 2020). Consider also those who argue for cognitive enhancement via external objects (e.g., Clark and Chalmers 1998; Kirsh 2010; Heersmink 2017).

possible to hold anyone else responsible" (Sparrow 2007, p. 65). Neither the programmers nor users (like the commanding officers), Sparrow argues, can be properly held responsible. Further, we cannot hold machines themselves responsible, given that they cannot appreciate such responses as punishment or reward. Indeed, even some techno-optimists are concerned that machines' inability to suffer thwarts our ability to directly hold them accountable—this is often why bridges are built indirectly, via the machines' associates.[15]

Yet, the problem runs deeper than our inability to exact punishment upon machines. Few will deny that technology has become increasingly ubiquitous. The harmful effects upon us are not just the overt attacks upon our bodily integrity, as from autonomous weapons or self-driving cars. To be sure, these sorts of overt harms are (hopefully) quite rare. Instead, what appears to add an especially persuasive fuel to the pessimist's agenda is that the effects of technology are often subtler but much more commonly encountered.[16] Consider the now widespread concern for undermined autonomy as a result of targeted advertising. Further, our devices are being made to perform tasks we aim to evade: remembering phone numbers or birthdays, navigating maps, or recipes. The problem is often referred to as AI *outsourcing*, wherein we effectively supplement—and perhaps degrade—our cognitive capacities, our professional skills, and even our interpersonal relationships (Vallor 2015; Danaher 2018, 2019). Can I really take credit, for example, for finding a great restaurant when I merely skimmed through Yelp? Is it just as meaningful to wish a childhood friend a "happy birthday" when an algorithm prompted me to do so? In short, just as technological entities themselves are presenting potential difficulties in locating responsibility, the technology-based gaps may be increasingly encountered in ourselves, the more we rely on technology for our leisure activities (like entertainment) as well as more basic needs (like healthcare and financial security). Bridging these gaps, then, requires not only wholescale reforms in personal awareness and international governance but also a candid reevaluation of who we are and where we are ultimately heading. Given that we are not likely to pursue these grand efforts, the pessimists can argue, we are better off without such disruptive technologies.

3 Shifting the Question

Undoubtedly, the burgeoning discussion, both in academic circles and in popular media, contains no shortage of reasons to be pessimistic about the present and future state of our moral well-being as a result of technology. The processes by which we understand and assign responsibility appear to be among our most fundamentally human qualities. As Strawson (1962) famously maintained, our engagement in interpersonal relationships, animated by the reactive attitudes, is something we cannot be without. Considering the increasing complexity of our environments and goals, Matthias claimed, our use of AI systems is now something "we cannot *do* without"

[15] For example, Nyholm (2018, 2020) accepts that robots cannot suffer and, nonetheless, argues that the collaborative agency found in human-robot partnerships can be plausible loci of responsibility.

[16] With the increasing prevalence of AI and robotic technologies in our day-to-day lives, and the potential effects upon our moral well-being, it seems there are reasons to support a process of co-development. See Tigard et al. (2020).

(Matthias 2004, p. 183, italics added). For those inclined to grant the plausibility of these outlooks, we should be quite concerned at the risks our technologies are presenting. We must not undersell nor should we overstate the potential conflicts. However, we should assure that there is a conflict in the first place. Putting aside the extent to which it is soluble, I want to step back and ask whether or not there is a uniquely technology-based responsibility gap.

In a paper posing the question of "Technologically blurred accountability?," Sebastian Köhler, Neil Roughley, and Hanno Sauer (2017) directly address the problem raised by Matthias (2004). Though they arrive at a comparable negative conclusion as drawn here, they do so in a way that is too narrow to fully escape concerns raised by Matthias and the pessimists. To grasp the limitations, we must recognize, specifically, their framing of responsibility gaps and, generally, their understanding of responsibility.

Köhler, Roughley, and Sauer aptly characterize the responsibility gap as a "normative mismatch." For Matthias and others, machine learning takes away the requisite control from designers, programmers, and users. When technologies capable of autonomous learning and action are causally involved in bringing about harm to humans, we simply cannot say who (or what) is responsible. Normally—and as traditional responsibility theories hold—we look for those with awareness and control of the action performed. According to tradition, that is, we *should* be able to locate responsible parties by locating those who knew and could have done otherwise. Unfortunately, if the pessimists are right, traditional accounts fail in cases of truly autonomous behavior performed by machines, which are of course not natural moral agents. While we should be able to locate responsibility, we lack the resources to do so, and here is where we see the mismatch. For Köhler, Roughley, and Sauer (2017), responsibility gaps have two key features:

> (1) it seems fitting to hold some person(s) to account for some φ to some degree D. Second, in such situations either (2.1) there is no candidate who it is fitting to hold to account for φ *or* (2.2) there are candidates who appear accountable for φ, but the *extent* to which it is, according to our everyday understanding, fitting to hold them individually to account does not match D. (p. 54)

No doubt, the framing of the responsibility gap as a normative mismatch provides refreshing clarity to the problem. Still, it is not clear why we must understand responsibility only in terms of accountability. As they acknowledge, and as a great deal of contemporary literature attests, accountability is merely one type of responsibility.

Although an extensive history cannot be given here, a brief background on pluralism in moral responsibility should suffice. Above, I introduced Strawson's notion of the reactive attitudes, which single-handedly moved theorists away from traditional accounts whereby we seek merely an agent's knowledge and control. Building upon our natural human reactions, contemporary theorists have recognized that this mechanism for locating responsibility is not a singular enterprise. To be sure, Köhler, Roughley, and Sauer (2017, p. 52) accept that there are "at least three different things that could be meant by saying that A is responsible for some φ." Following Gary Watson and David Shoemaker, it is pointed out that responsibility can be a matter of attributability, accountability, or answerability. For Watson (2004), the first two "faces" of responsibility come apart with the observation that we make judgments of others' practical identity, what they value, or stand for. *Additionally*, we might—or might not—hold them to account, namely by communicating our offense or punishing the offender. Similarly, for Shoemaker (2011, 2015), to attribute an action to someone is to think it reflects their underlying cares or commitments, whereas to hold one to account is to engage in more overt forms of blame, like directed anger. Aside from these two sorts of responses, however, we might also demand answers, particularly for the harms that befall us. Answerability, on Shoemaker's account, is a process by which we call upon others to provide explanations, in order to evaluate their judgment and understand their reasons for action.

According to Köhler, Roughley, and Sauer (2017), "the worry about responsibility becoming blurred is mostly about *accountability*" (p. 52, italics in original). But this appears to be the case only if our focus is restricted to current regulatory gaps. As I have maintained, moral responsibility is not the same as accountability via regulation—and Matthias (2004) explicitly recognized this by introducing his discussion with reference to Strawson's reactive attitudes and theories of moral responsibility (Fischer and Ravizza 1998; Oshana 2002).[17] My suggestion here is that just as we may try to locate accountability in technology, we may—and I believe we *do*—at least attempt to engage in other forms of responsibility practices. I will have more to say on the details of these processes in the following section. For the moment, we need only to reflect on the variety of ways in which we now interact with technology to see that our concerns go beyond accountability. For example, we might think of technology as reflecting moral norms or values—indeed, we have long seen appreciation of the idea that technology is not value-free (Friedman 1997).[18] Likewise, we might try to understand reasons for the decisions and behaviors exhibited by technological entities, perhaps in order to maintain a degree of human participation.[19]

This is not to say that our attempts to pinpoint values in technology or to understand reasons for its behavior will succeed. Undoubtedly, such attempts often fail and will likely continue to do so. The point to be made here is that these efforts represent attempts to locate responsibility in technology and, importantly, the sorts of responsibility in question cannot be characterized as accountability. In this way, if we limit the focus to accountability, we leave open the possibility of technology-based gaps arising from the lack of attributability or answerability in technological entities. Notice, also, that our diverse responsibility practices can take on both forward- and backward-looking directions, as it were. In other words, we hold others responsible not only for what *has happened*, but also for what we demand and would like to have happen in the future. On the account of Köhler, Roughley, and Sauer (2017), the sort of

[17] Note that admitting moral responsibility to be a pluralistic enterprise would appear to open up additional gaps in responsibility. Indeed, I see this line as a potential challenge. However, in this way, addressing the plurality should provide a stronger argument—that is, against any sort of gap—than those focused only on accountability.

[18] Here I should note that I remain neutral on the issue of values being substantively embedded in technology or remaining only in the users—what Friedman (1997) called the endogenous versus exogenous views, respectively. My arguments do not rely on either and should be amenable to proponents of both ends as well as those in the middle, maintaining (say) that technology can help to enable or disable particular human actions.

[19] Danaher (2016a) aptly argues that because our increasing reliance upon algorithms limits human participation and comprehension, algorithmic systems pose a threat to legitimate public decision-making.

"responsibility in question is *retrospective*, rather than *prospective*" but it is unclear why (p. 52, italics in original).

In sum, limiting the question of techno-responsibility gaps to retrospective accountability does not fully satisfy the broader inquiry into whether or not there is a technology-based gap to begin with. For those seeking values or reasons in technology, and for those demanding certain outcomes for the future, decisive questions are left unanswered by explaining away only cases of blurred accountability for the past. Can we, or can we not, *attribute* some action to an AI system? Is there a sense in which we can hold technological entities *answerable* for their conduct? Köhler, Roughley, and Sauer help us to see that the supposed technology-based gaps in *accountability* result from epistemic or pragmatic challenges. It may be, for example, that one cannot foresee the behavior of autonomous devices, or that it will be too difficult to trace back a causal connection to anyone in control of the outcome. All the same, they argue, we can assign accountability to those who risk unforeseen harm or to those beyond direct control of a subordinate agent—like dog owners held to account for damages caused by their dogs (Köhler et al. 2017, pp. 58, 63). Yet, for those with concerns beyond punishment for past harms, the problem of potential technology-based gaps in responsibility remains open. In the following section, I aim to resolve this wider issue by showing that responsibility, with its diversity of forms and functions, is robust enough to encompass emerging technologies.

4 Demanding Answers and Understanding Values

Allow me to briefly take stock of my overarching argument thus far. First, AI and autonomous technologies seem to pose an extraordinary threat to responsibility—a premise accepted by those who argue against deploying such systems, as well as those who argue that the gap can be bridged. Few have made the case that there is no techno-responsibility gap, but those who have (Köhler et al. 2017) focus only on accountability for the past. Yet, responsibility is a broad enterprise, as I have suggested.[20] Thus, in order to show that technology is truly not a threat to responsibility, it must be shown that there is also no technology-based gap in answerability, attributability, and forward-looking accountability. It is to this task I now turn.

Consider that there are cases wherein the principal cause of harm is a technological entity, such as an autonomous weapon misfiring upon innocent civilians or an algorithm trained by a social media user to display increasingly violent content. For Köhler, Roughley, and Sauer, such situations are not necessarily a threat to our "ordinary conception" of accountability. As they demonstrate, we can nonetheless hold someone accountable, namely those who risked harm (perhaps negligently), or all of those who made some minimal causal contribution, say, by assisting in the design or programming of software. In short, while (1) it seems fitting to hold some person to account, most often we have no trouble doing so—that is, we need not affirm that (2.1) there is no candidate who it is fitting to hold to account.

[20] Although I have not offered a unique argument for responsibility pluralism, I can only hope that readers will have at least a hunch that responsibility is more than accountability. For those not convinced, I again draw attention to the theories defended by Watson (2004) and Shoemaker (2011, 2015).

By looking beyond accountability, we see that some who are harmed by technology will want to better understand the reasons for a machine's behavior or the underlying values that it seems to have been programmed or learned to promote.[21] In other words, cases will be encountered wherein (1*) it seems fitting to demand answers or to attribute the conduct to some underlying set of values or commitments, but also where (2*) there is no fitting candidate from whom answers can be demanded or to whom the conduct can be attributed. Here we have two additional normative mismatches, based respectively upon the demands for answerability and attributability. Accordingly, even where we grant that cases of technology-based harms are not threatening to our ordinary conception of accountability, we see other potential sources of a technology-based gap. We must, then, account for this variety of ways in which we hold one another responsible. If it can be shown that 2* is deniable, or that 1* need not be affirmed, then the additional normative mismatches are unproblematic and it would appear there are no techno-responsibility gaps, in a much broader sense.

First, consider answerability, which Shoemaker introduces as evaluating one's judgment. It may seem that in order to demand answers for a machine's behavior, we must call upon the designers, programmers, or users.[22] Yet, here we fall back into Matthias's responsibility gap, as one of the key problems with AI technology is that it will learn on its own—with artificial neural networks, decisions, and patterns of behavior are adopted in ways we cannot understand (Matthias 2004, pp. 179–181). Designers, programmers, and users of some devices, very often, will be unable to *answer for* them. As a result, we can hold these others accountable—say, with legal ramifications or social stigma for tech companies—but we cannot sensibly demand answers. Those associated with the production and use might simply not know why an AI system exhibited some behavioral pattern rather than another. Why, then, would we not demand answers from the system itself?

Intuitively, it seems odd to think machines could form judgments resembling the judgments of natural moral agents. They cannot be said to truly entertain or act upon reasons.[23] But notice that something like the process of demanding and receiving answers can take place in our interactions with AI. Consider the increasingly intelligent sensory capacities of many AI systems, like autonomous vehicles equipped with an array of highly sophisticated cameras. Many of today's technologies are able to retain sets of data that easily surpass human cognition and are able to learn about their environments in details we can only imagine. While datasets are not *reasons* as we

[21] Again, I am not saying that values can be substantively embedded—or that they cannot be. The point is simply that some who are affected by technology might try to understand values that are somehow reflected (whether embedded in devices or residing ultimately in designers and users). Similarly, as I reiterate below, whether or not machines can act *upon* reasons, some will want to grasp reasons for their behavior, at least in a general sense.

[22] Coeckelbergh (2019) takes this approach. And while the paper highlights the importance of "answerability" (and relational responsibility, generally), it does so without accounting for the notion as it has been developed in the responsibility literature. Coeckelbergh (2019, p. 13) also maintains a focus on "*human beings* who are able to explain things," but thereby creates a (minimally) 3-way relation, obscuring the agent-patient relation he aims to illuminate.

[23] Some who maintain this view (Purves et al. 2015; Talbot et al. 2017) do so on the grounds that machines lack the relevant kind of mental states, namely phenomenal consciousness. By contrast, Frank and Nyholm (2017) adopt a functionalist view of agency and argue that robots could act on the basis of reasons. I appreciate an anonymous reviewer pointing out this relevant debate, and without committing myself to either view, I assume only that technological devices cannot act upon reasons in the fullest sense, namely as we can.

know them, AI systems can obtain and analyze a myriad of environmental features in ways that issue in calculated decisions and distinct behavioral outputs. Where this behavior is unrecognizable, or perhaps exceeds the purposes of initial programming, we can engage in efforts to decipher reasons. We can work to understand a system's initial programming, any newly adopted algorithm, and the relevant set of inputs, such as the data used to train machine-learning systems.[24] Granted, again, our efforts might not succeed—some AI systems may well remain "black boxes." However, fully functional adult human beings also fail to provide adequate reasons for their decisions, say, due to implicit biases (Doris 2015; Vargas 2017) and this has not stopped us from demanding answers from them. What is important here is that as a social or political process, one with great moral significance, we hold technological systems answerable, even if doing so requires the aid of human experts.[25] Our investigations might fail to produce adequate reasons for questionable outputs, but this is not to say we cannot appropriately make such demands. Further, whether or not satisfactory answers are received, by engaging in answerability practices, we can be better prepared to instigate additional responsibility practices, such as holding rogue systems and any human associates to account, as I explain in more detail below.[26]

Turning to attributability, which evaluates one's character, again it would appear most plausible to seek this sort of responsibility in any humans associated with the outcomes brought about by technology. After all, attributing some action or attitude, for Watson and Shoemaker, requires one to be capable of maintaining cares and commitments, of displaying the values that one *stands for*. Yet, again, any human associates—regardless of their values and commitments, and how well those may have been instantiated in technology—are effectively dissociated once a machine learns and behaves in ways that no longer reflect the initially programmed values and commitments. Thus, we fall back again into the techno-responsibility gap, for we cannot locate a source of responsibility in terms of attributability. Further, looking to the device itself offers only very limited help here.

Even granting that technology can reflect human values—or that AI can possess something like "functional morality" (Allen and Wallach 2009, 2011)—surely the most intelligent machines cannot truly be said to *care*. They cannot be committed to something or someone in ways that allow us to assess *who they are* or *what they are like*. This manner of evaluation seems reserved exclusively for natural moral agents, namely ourselves. Attributing behavior to the "character" of a non-human animal or machine is only metaphorical or symbolic. That being said, we commonly hold non-human animals responsible in terms of attributability, at least very loosely, when we think their conduct expresses something about them. Consider those who believe their dog to be a hero for saving a drowning child. Similarly, we might blame our devices for

[24] In fact, Matthias himself floated this method of indirect explanation, suggesting that the "knowledge and behavior stored in a neural network can be only inferred indirectly through experimentation and the application of test patterns after the training of the network is finished" (2004, p. 181).

[25] A problem that arises here is that we have simply traded the potential techno-responsibility gap for an overreliance upon an epistemic elite (see Danaher 2016a, p. 255). Nonetheless, the latter is clearly a quarrel among humans, which can in principle be addressed by regulatory measures.

[26] It should also be noted that in cases where systems are utterly incomprehensible, even to relevant experts, we likely cannot affirm 1*—the claim that it is fitting to demand answers. In this way, there is no gap. I employ this line of argument below, in response to concerns for attributability.

possessing unique peculiarities—picture the student who claims "*My* laptop freezes whenever I try to print!" No doubt, attributing an action or attitude to the "character" of a non-human animal or a machine is far from natural and largely incoherent, however comical. Notice, however, that where an attribution of character to technology appears incoherent, it will appear similarly unfitting to seek an underlying set of values or commitments. In other words, while it is difficult to deny 2*—the claim that there is no fitting candidate to whom the conduct can be attributed—claim 1* (the fittingness of attributing the conduct) was not affirmed.[27] For this reason, when it comes to attributability, there is not a normative mismatch and, thereby, no techno-responsibility gap. Still, I claimed, some who are harmed as a result of technology will seek to understand the underlying values embedded. In this way, to the extent that it is seen as fitting to attribute some conduct to an underlying set of values, one will likely be comfortable denying that there is no fitting candidate. That is, the designer, user, or perhaps the device itself will be seen as reflecting some perceived value.

Finally, although Köhler, Roughley, and Sauer work to repair the normative mismatch in accountability, this effort is too narrow. As stated, their focus is solely on locating accountability for the past. Yet, clearly, holding others to account can be forward-looking in its nature—we express anger at friends who mistreat us, and we lock up violent criminals, often to prevent similar harms in the future and perhaps not because we think the target of our response *deserves* it.[28] In fact, by focusing only on retrospective accountability, we risk playing into the techno-pessimists' fears. Recall Sparrow and others concerned at our inability to punish AI due to machines' inability to suffer. Köhler, Roughley, and Sauer will respond by suggesting we hold to account those who risked harm, such as commanding officers deploying autonomous weapons. But here, pessimists can retort that this mechanism fails to repair the mismatch, namely because it is not entirely fitting—nor as satisfying psychologically—to blame *others* for some harm (Danaher 2016b). Worse, consider again that we fall into responsibility gaps when maintaining natural reactions toward absent, unknown, or nonexistent sources. Even if blaming AI's associates is successful, there will be cases where there are no associates to blame. Once again, the question becomes, can we locate this sort of responsibility in technology itself?

If accountability is only retrospective, it is difficult to understand what it would mean to hold AI to account. After all, for many responsibility theorists, holding others accountable is essentially communicative—anger or resentment are means of eliciting guilt and reparative actions in the offender. What gives effect to such communication is the offender's capacity to empathize (Shoemaker 2015, p. 111). Here we see why some are concerned for newfound instantiations of psychopathy in heartless AI systems (Coeckelbergh 2010). Fortunately, our responsibility practices often take on prospective aims. We hold others to account by communicating wrongdoing and even punishing, but not necessarily because they deserve to suffer. Very often, we want to see that the future is better than the past. We want to help others learn from their mistakes and discourage repeated harmful behavior. Undoubtedly, AI systems are

[27] Although he does not distinguish between the various sorts of responsibility, a similar move is made by Kraaijeveld (2019).

[28] Indeed, there are those—like Pereboom (2014)—who adopt a purely forward-looking account of responsibility.

unable to suffer like us, but they can and often should be targeted with reparative measures. Like our accountability practices toward fellow humans, we can hold AI to account by imposing sanctions, correcting undesirable behavioral patterns acquired, and generally seeing that the target of our response works to improve for the future—a bottom-up process of reinforcement learning (Allen and Wallach 2009; Hellström 2013). In short, the potential normative mismatch in accountability can be repaired indirectly by retrospective measures toward human associates (designers, programmers, or users) or directly by adopting forward-looking aims in our interactions with technology.[29]

Recall my admission that in cases like the self-driving lorry—where AI systems cannot hold themselves responsible—it appears we face a techno-responsibility gap. Before we discover that the source of harm is only a machine, I claimed, we may well respond naturally, as if it were a moral agent like us. Of course, AI systems are not full-fledged moral agents, and they may well never display good or ill will, but only indifference, in Strawson's words.[30] Yet, this fact should be instructive and not discouraging to our efforts at holding them responsible. Once we learn that the sources of harm are purely technological entities, certainly, we modify our attitudes and practices. But the modifications taking place are not like the excuses we grant to fellow humans in extraordinary circumstances. Instead, like our interactions with children and psychopaths (among others), we realize that the potential target of our moral attitudes is exempt from our *usual* responsibility practices. This is not to say we do not deploy responses of *any* sort. Rather, we adjust to distinct agential conditions, which may well allow us to discern reasons, understand underlying values, or even communicate in ways that improve the future.[31]

5 Conclusion

Following Strawson's naturalistic turn away from traditional notions of responsibility, the pluralistic accounts of Watson and Shoemaker have helped us to see that responsibility is a dynamic enterprise wherein we interact with an immense range of individuals, some of which at the far reaches of our moral universe. Often, as Shoemaker emphasizes, we respond with ambivalence, indicating that we hold some individuals "responsible in some ways but not in others" (2015, p. 3). For those we exempt from

[29] Recent work supports the idea that AI systems will become increasingly able to recognize and learn from our morally significant reactions. See, e.g., Ren (2009) and Knight (2016).

[30] Again, I thank an anonymous reviewer for encouraging me to clarify this point.

[31] Here one could envision a table, which I will only describe, since the same kind is seen in Shoemaker (2015, pp. 123–124)—namely where each "marginal" agential condition in humans (psychopathy, autism, etc.) is listed vertically and, across, a status is given for each type of responsibility. For instance, dementia is summarized as follows: attributability (unmitigated); answerability (exempt to mitigated); accountability (exempt to mitigated). Following this rubric, my account of sophisticated machine-learning systems can be summarized as follows: answerability (mitigated, depending upon the extent to which we discern reasons); attributability (mitigated, depending upon the extent to which we understand values); accountability (retrospective: exempt unless indirect; prospective: unmitigated). Notice that—for both Shoemaker's account of human conditions and my account of technological systems—for some responsibility types to be mitigated, or even exempt, does not mean we do not hold those agents responsible *full stop*. Instead, we see indications of our ambivalence, in the sense that we are inclined to hold them responsible in some ways but not in others.

the full force of our reactive attitudes, in Strawson's terms, we adopt an "objective" stance rather than maintaining the "participant" attitudes expressive of our interpersonal relationships. As Watson notes, even when adopting the objective view of others, we see them as individuals "to be controlled, managed, manipulated, trained" (2004, p. 225). AI systems too are naturally exempted from our usual moral attitudes, but they can nonetheless be controlled, managed, manipulated, and trained.

By taking seriously the variety of ways in which we hold others responsible, we see that the threat to responsibility posed by emerging technologies is far less severe than most authors have conveyed. As I showed, those aiming to bridge the gap, as well as those calling for a moratorium on AI development, both accept that there is problematic techno-responsibility gap. And while it may appear that I have lent more support to the techno-optimists—those who wish to proceed with AI in domains like medicine, self-driving cars, or warfare—I want to reiterate that my goal has been to clarify a different question. Before asking how, if at all, the gap can be bridged, we must first ask whether or not there is a uniquely technology-based responsibility gap. Importantly, this latter inquiry goes beyond accountability for past harms. Along with locating accountability, we can demand answers and work to better understand the values being promoted by designers, users, and by technology itself.

Acknowledgments This paper began as a follow-up to my "Artificial Moral Responsibility" paper and as a contribution to David Lawrence and Sarah Morley's *Novel Beings* project. I am grateful for their exceptional symposia held at Newcastle University in April 2018 and the Wellcome Collection in October 2018. I also benefitted from discussions at RWTH Aachen University, where I received postdoctoral support, and the Freiburg Institute for Advanced Studies, with colleagues including Niël Conradie, Philipp Kellmeyer, Peter Königs, Carmen Krämer, Saskia Nagel, and Sven Nyholm. For comments on earlier drafts, a debt of gratitude is owed to Nathan Biebel, Peter Königs, and my anonymous reviewers.

Funding Information Open Access funding provided by Projekt DEAL.

References

Allen, C., & Wallach, W. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Allen, C., & Wallach, W. (2011). Moral machines: Contradiction in terms or abdication of human responsibility? In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 55–68). Cambridge: MIT Press.

Archard, D. (2013). Dirty hands and the complicity of the democratic public. *Ethical Theory and Moral Practice, 16*(4), 777–790.

Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*, 687–709.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in healthcare–Addressing ethical challenges. *New England Journal of Medicine, 378*, 981–983.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, forthcoming. https://doi.org/10.1007/s11948-019-00146-8.

D'Arms, J., & Jacobson, D. (2000). Sentiment and value. *Ethics, 110*(4), 722–748.

D'Arms, J., & Jacobson, D. (2006). Anthropocentric constraints on human value. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics, vol. 1* (pp. 99–126). Oxford University Press.

Danaher, J. (2016a). The threat of algocracy: Reality, resistance and accommodation. *Philosophy and Technology, 29*(3), 245–268.

Danaher, J. (2016b). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309.

Danaher, J. (2018). Toward an ethics of AI assistants: An initial framework. *Philosophy and Technology, 31*(4), 629–653.

Danaher, J. (2019). *Automation and utopia: Human flourishing in a world without work*. Harvard University Press.

de Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics, 26*(2), 727–735.

Dignam, A. (2020). Artificial intelligence, tech corporate governance and the public interest regulatory response. *Cambridge Journal of Regions, Economy and Society, 13*(1), 37–54.

Doris, J. (2015). *Talking to our selves: Reflection, ignorance, and agency*. Oxford University Press.

Fischer, J. M., & Ravizza, M. S. J. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial Intelligence and Law, 25*(3), 305–323.

Friedman, B. (1997). *Human values and the design of computer technology*. Cambridge University Press.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91–99.

Heersmink, R. (2017). Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences, 16*, 17–32.

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107.

Jacobson, D. (2013). Regret, agency, and error. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility, vol. 1* (pp. 95–125). Oxford University Press.

Kirsh, D. (2010). Thinking with external representations. *AI & SOCIETY, 25*, 441–454.

Knight, W. (2016). Amazon working on making Alexa recognize your emotions. *MIT Technology Review*.

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability? In C. Ulbert et al. (Eds.), *Moral agency and the politics of responsibility*. London: Routledge.

Kraaijeveld, S. (2019). Debunking (the) retribution (gap). *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00148-6.

Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics, 6*(12), 46–51.

Mason, E. (2019). Between strict liability and blameworthy quality of will: Taking responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility, vol. 6* (pp. 241–264). Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Morley, S. (manuscript). Morally significant technology: A case against corporate self-regulation.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics, 24*(4), 1201–1219.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

Oakley, J. (1992). *Morality and the emotions*. London: Routledge.

Oshana, M. (2002). The misguided marriage of responsibility and autonomy. *The Journal of Ethics, 6*(3), 261–280.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14.

Ren, F. (2009). Affective information processing and recognizing human emotion. *Electronic Notes in Theoretical Computer Science, 225*, 39–50.

Sharkey, N. (2010). Saying "no!" to lethal autonomous targeting. *Journal of Military Ethics, 9*(4), 369–383.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*(3), 602–632.

Shoemaker, D. (2015). *Responsibility from the margins*. Oxford University Press.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Stout, N. (manuscript). Blame *de re* and *de dicto*.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.

Szigeti, A. (2015). Sentimentalism and moral dilemmas. *Dialectica, 69*(1), 1–22.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Talbot, B., Jenkins, R., & Purves, D. (2017). When robots should do the wrong thing. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 258–273). Oxford University Press.

Tigard, D. (2019a). Moral distress as a symptom of dirty hands. *Res Publica*, 25(3), 353–371.

Tigard, D. (2019b). Taking the blame: Appropriate responses to medical error. *Journal of Medical Ethics, 45*(2), 101–105.

Tigard, D. (2020). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*, forthcoming.

Tigard, D., Conradie, N. H., & Nagel, S. K. (2020). Socially responsive technologies: Toward a co-developmental path. *AI & Society*, forthcoming. https://doi.org/10.1007/s00146-020-00982-4.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy and Technology, 28*(1), 107–124.

van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67.

Vargas, M. (2017). Implicit bias, responsibility, and moral ecology. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility, vol. 4* (pp. 219–247). Oxford University Press.

Verbeek, P. P. (2008). Obstetric ultrasound and the technological mediation of morality. *Human Studies, 31*(1), 11–26.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review, 2019*(2), 494–620.

Watson, G. (2004). *Agency and answerability*. Oxford University Press.

Williams, B. (1981). *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.