Is AI a Problem for Forward Looking Moral Responsibility? The Problem Followed by a Solution

Fabio Tollon(✉)

Department of Philosophy/GRK 2073 "Integrating Ethics and Epistemology of Scientific Research", Bielefeld University, Bielefeld, Germany
fabio.tollon@uni-bielefeld.de

Abstract. Recent work in AI ethics has come to bear on questions of responsibility. Specifically, questions of whether the nature of AI-based systems render various notions of responsibility inappropriate. While substantial attention has been given to backward-looking senses of responsibility, there has been little consideration of forward-looking senses of responsibility. This paper aims to plug this gap, and will concern itself with responsibility as moral obligation, a particular kind of forward-looking sense of responsibility. Responsibility as moral obligation is predicated on the idea that agents have at least some degree of control over the kinds of systems they create and deploy. AI systems, by virtue of their ability to learn from experience once deployed, and their often experimental nature, may therefore pose a significant challenge to forward-looking responsibility. Such systems might not be able to have their course altered, and so even if their initial programming determines their goals, the means by which they achieve these goals may be outside the control of human operators. In cases such as this, we might say that there is a gap in moral obligation. However, in this paper, I argue that there are no "gaps" in responsibility as moral obligation, as this question comes to bear on AI systems. I support this conclusion by focusing on the nature of risks when developing technology, and by showing that technological assessment is not only about the consequences that a specific technology might have. Technological assessment is more than merely consequentialist, and should also include a hermeneutic component, which looks at the societal meaning of the system. Therefore, while it may be true that the creators of AI systems might not be able to fully appreciate what the consequences of their systems might be, this does not undermine or render improper their responsibility as moral obligation.

Keywords: Responsibility gaps · Forward-looking responsibility · Technological assessment · Moral obligation

1 Responsibility and AI

Questions of responsibility have become increasingly important in the field of AI ethics. Specifically, questions of whether the nature of certain technological systems equipped with AI renders various notions of responsibility inappropriate. Substantial attention has

E. Jembere et al. (Eds.): SACAIR 2021, CCIS 1551, pp. 307–318, 2022.
https://doi.org/10.1007/978-3-030-95070-5_20

been given to backward-looking or retrospective senses of responsibility, as these come to bear on questions of moral responsibility [12, 15, 17, 20]. However, less attention has been paid to *forward-looking* senses of responsibility. This paper aims to plug this gap, and so I will concern myself with responsibility as moral obligation, a particular kind of forward-looking sense of responsibility.

Responsibility as moral obligation is a responsibility for *future* states of affairs and is concerned with the active promotion of certain societal goals, and the responsibility of agents to align what they do with these goals [17]. We must take seriously our obligation to ensure that the decisions we make today help in the pursuit of a better tomorrow. This is not to merely suggest that we take the future into account, but rather that we have an *active* obligation to steer society in a way that aligns with various important values.

For an agent to be responsible for future states of affairs, the means by which they go about getting to these future states ought to be in some sense under their control, such that if a desirable outcome is not achieved, we would be able to find them at fault. For the agent to be *responsible*, in this sense, it needs to be the case that the desired future state of affairs is somehow *up to them.* That is, it is *possible* for them to see to it that the future state comes into being. Thus, this moral obligation is predicated on the idea that agents at least have some degree of *control* over the kinds of systems they create and deploy. AI-systems, by virtue of their ability to learn from experience once deployed, and their often-experimental nature, may pose a significant challenge to responsibility as moral obligation. Such systems might not be able to have their course altered, and so even if their initial programming determines their *goals,* the means by which they achieve these goals may be outside the control of human operators. In cases such as this, we might say there is a gap in moral obligation. It is this "gap" that this paper aims to plug.

2 Task, Authority, and Obligation

Engineers and the creators of technology have a moral obligation to ensure that their products comply with certain norms and standards, and that these are in the service of socially desirable values and goals. This is not news to those working in engineering ethics and related disciplines. That engineers have a responsibility to design, for example, bridges that can bear a certain weight, buildings that can withstand certain wind speeds, and transportation systems that do not endanger passengers is nothing new or controversial. However, these obligations are closely linked with the *tasks* that engineers conduct. That is, they have a professional obligation to design and build structures that adhere to certain basic requirements. Moreover, they are also *authorities* on such matters (given their education) and are responsible for how the project is in fact carried out. This kind of responsibility is termed role responsibility and is *descriptive* in nature. It refers to whether the agent in question is in the correct kind of causal relation to an outcome, given their position in an organization or their authority. We can therefore think of this descriptive sense of responsibility as being *passive* in that it is more concerned with the legal and moral consequences that might be brought to bear on engineers should they fail to perform their task *as engineers.*

In addition to this, we have responsibility as *moral obligation*, which differs from responsibility as task and authority in that it is not so closely coupled with the *technical*

skills of the engineer but is a *normative* sense of responsibility. It is also focused on responsibility for future states of affairs and is therefore not restricted to the analysis of an agent's adherence to certain professional norms. It is therefore a predominantly forward-looking sense of responsibility. We are here concerned with a *prescription* that agents have, "in terms of an obligation to do something or to see to something, or to take care of something" [16]. Here the question is not necessarily about what responsibilities they may have given their position as engineers, but rather on the responsibilities that we might expect them to rationally or reasonably assume [16]. The responsibilities we might expect them to assume are natural extensions of their roles and authority, but they are distinct in that when it comes to moral obligation there is an *active* component: they are expected to check-in, supervise, and *ensure* that some future state of affairs obtains. It is therefore important that we keep the distinction between "role responsibility", on the one hand, and moral responsibility on the other. Role responsibility has to do with certain professional norms (as noted earlier) while moral responsibility, in the forward-looking sense, has to do with responsibility for future states of affairs.

A second thing to note is that it is not controversial that these forward-looking obligations change over time: for example, added to the list of forward-looking responsibilities that current engineers have is a concern for sustainability, which may not have been a factor just 80 years ago. Nor would such a change (necessarily) lead to a "gap". What I will now investigate is whether AI systems might come to complicate our ability to make evaluations of forward-looking responsibility due to their unpredictability and their experimental nature.

3 Passive and Active Responsibility

As mentioned previously, responsibility as moral obligation is concerned with seeing to it that something is the case. For an agent to be reasonably held responsible in this sense, we expect that they could, for example, check in on the system or update it, so that the desired future state of affairs is achieved. It is their ability to intervene in this way that makes them fairly responsible. This, however, becomes difficult in the case of those developing AI-systems. AI research is an innovative and unpredictable field, due to both the vast collection of agents involved in creating these systems and the nature of AI itself. In the first case, engineers may have competing obligations towards different stakeholders (their employers, the public, etc.), and they might not be aware of what their social roles entail (whether they are scientists, businesspersons, or technicians, etc.). In the second case, the nature of AI-driven systems makes them in some sense *experimental*, in that engineers are often testing and *innovating* with potential solutions, without knowing exactly what the future consequences of their decisions may be. Additionally, once these systems are deployed, it may be impossible for engineers to actively intervene: the system would be *outside of their control*.

A gap in moral obligation may occur when we reflect on these two features. In the first case, engineers may not be aware of their obligations to other agents (due to the complicated network in which they are embedded) and therefore cannot seemingly be held responsible for a failure to meet these obligations. However, this is not unique to AI development: there are other contexts in which agents are embedded in complicated

bureaucratic networks and may be unsure of their obligations due to this. For example, the scientists working on the Manhattan Project, which produced the world's first nuclear weapons during World War II, worked in massive teams of discrete groups. This was the beginning of Big Science, where scientists were not limited to working in siloed academic laboratories but were pushed into large-scale organizations with political backing (and political agendas) [18]. This kind of project was massively funded and had clearly defined objectives. The important point here is that this represented a break in the "traditional" way of doing science and created a bridge between the power of science and political power. The mingling of these two once broadly distinct spheres resulted in a complication of the scientists' understanding of their own moral obligations, as they were not "married to the science" as it were, but also to various political agendas, which they might depend on for future funding, employment, access to resources, etc. My point is not to go into the details of this affair, but rather to note that it created an organizational structure in which the role that scientists found themselves to be playing was not only restricted to science itself. As Steven Shapin put it,

> Scientists had never before possessed such authority, largesse, civic responsibility, and obligations. By free choice or not, some scientists now lived the *vita activa*, and, while there were still consequential worries about the extent to which they were indeed "normal citizens," they had never been more integrated into the civic sphere [18].

Here we can see how scientists' obligations *qua* scientists might be complicated by their new political power. This could be thought to create a gap in moral obligation, due to the often competing (or conflicting) objectives of scientific research and political objectives. If scientists are not properly trained for their new political roles, we might expect them to be unable to cope with these new obligations. It may therefore be unreasonable to expect them to fully understand their roles, especially in the early days of this mixing of political and scientific power. They might not understand their *active* responsibilities with respect to their research and how it is implemented. The point is that such complications of responsibility as moral obligation, at least in the sense of agents being answerable to different stakeholders, are not unique to AI.

Thus, the advent of AI and the teams of computer scientists behind them do not necessarily create a *unique* gap in moral obligation when we reflect on the issue of competing obligations. This does not mean the problem is solved, but rather that the question of competing stakeholders and the ways in which this challenges our ascriptions of responsibility cover much broader terrain than my present concern. Where AI might indeed pose a unique challenge, however, comes about in how it complicates the second feature of responsibility as moral obligation: the innovative and experimental nature of AI research and development may undermine the relevant *control* required for reasonable ascriptions of forward-looking responsibility. There are two aspects to this. The first concerns the *risks* that are inherent in deploying autonomous systems that are outside of human control. The second concerns our inability to predict the consequences of this technology.

4 Risks and Consequences

It seems it would be unreasonable to hold engineers responsible when they are dealing with experimental and innovative technologies. In such cases where full knowledge of the consequences is impossible to ascertain beforehand and intervention once the system is deployed is not possible, ascriptions of full responsibility do not seem to be fair. In the case of nuclear weapons and the Manhattan Project, it was relatively clear that this technology posed an existential threat to humanity should it ever be developed and deployed. While this was not a widely held belief among individual scientists at that time, there were many (such as Einstein) who were cognizant of this threat. One reason for this is the technology had a clearly defined purpose, and once developed, would be used in a very specific way. It would be deployed at targeted locations, determined by human agents. Up until the moment the bomb was dropped, there existed the possibility of human intervention, and at no point was it outside of meaningful human control. There was no sense in which the technology could set its own goals or learn from its "experiences". For the scientists (and all those involved in the project) there was therefore a sense in which they could reasonably *anticipate* the future *consequences* of what they were producing. This was true in two senses. In the first case, it was possible to anticipate the kinds of *risks* associated with the technology. Second, due to the direct nature of how the technology was to be deployed (and the nature of how the technology was tested beforehand), they could also reliably predict what the consequences of its use would be (in the sense of anticipating the level of destruction, not the effects on society overall). Obviously, these two factors are linked, but I think it is useful to consider them separately, as there may be cases where we understand the risks of a technology but do not have a clear handle on its potential consequences.

This, I think, might be what motivates the emergence of a gap in moral obligation due to AI: even if we grasp the risks, this does not necessarily entail that we can fully appreciate the potential consequences, which are often compounded by our lack of control over the system. I will argue, however, that both of these aspects can be overcome, and that there is therefore no gap in obligation. To do this I will first show that we do in fact, broadly, understand the risks of deploying AI systems (or at the very least, that we *know* about these risks), and we have frameworks in place to mitigate such risks. This will involve taking seriously that the risk of deploying a system over which we have no control might itself be a problem. Releasing fully autonomous AI out into the world is a choice we make, and we might be better off ensuring that we *always* have a sufficient level of control over such systems. Second, I will show that while an exclusive focus on consequences might allow for the emergence of a gap in responsibility, this is not the only means we have at our disposal in our assessment of technological systems.

5 Risks of AI

From the perspective of forward-looking responsibility, the question is not whether this makes blame appropriate for these systems, but rather whether such deployment creates a gap whereby developers, engineers, programmers, etc. cannot fulfil their future-orientated obligations because the AI-system has functional autonomy. That is, it can

operate outside of "meaningful human control" [13]. The main concern here is the supposed fact that it would be impossible for agents to intervene with such autonomous systems, foreclosing assignments of prospective responsibility. However, this claim operates with the assumption that AI somehow has a pre-configured position in society. I will argue, however, that AI systems, while importantly different to traditional technological artefacts, are nonetheless still created and deployed by human beings, and so their position in society is always a *choice* that *we* make [5].

While it is important to note AI is different from traditional technical artefacts, the one way in which it is similar is that it is *designed by and for human agents*. When thinking about AI we should therefore not be misled by the supposed "intelligence" of such systems. Their intelligence, if they have any, is of a derivative kind, and is the product of a *human process* of research and development. For example, Joanna Bryson argues that concepts such as intentionality, consciousness, and sentience, etc. are mere sideshows to the real problems posed by AI: these being problems of *governance* [6]. Specifically, she argues that the most pressing question concerning AI is how to design our artefacts "in a way that helps us maintain enough social order so that we can sustain human dignity and flourishing" [6]. From this perspective, questions of a forward-looking gap in moral obligation do not seem to arise. While of course, the potential agential nature of AI in many contexts makes it more *difficult* and can increase the *complexity* of determining what our forward-looking obligations might be, and what the best route to achieve them might entail, this does not by itself create a *gap*. For there to be real indeterminacy, the technology would seemingly have to come out of the ether with its own set of values and goals, in which case we would have no understanding of, or contribution to, its design. However, notice that even in this extreme case, the notion of forward-looking responsibility still makes sense: even if no human agent contributed to the system, we might still *reasonably expect* those with the relevant skills and competencies to intervene, check on the system, try to stop it, etc. Although we might not *blame* them for not doing so, this does not mean that they have failed in their active responsibility.

While it is of course true that the emergence of potentially autonomous systems requires critical ethical reflection, we as human beings still have full control over *when* and *how* AI systems are developed, and thus carry the responsibility for them [5]. This is especially true if we conceive of forward-looking responsibility as being concerned not only with desirable outcomes but also with the promotion of shared values. Here we would be interested with questions regarding the *alignment* of AI research and deployment with certain values that are deemed beneficial for society [4, 7]. As more and more social processes become automated and "outsourced" to AI or algorithmic systems, we need to understand these systems as not merely technical but rather as social and political artefacts, capable of reinscribing and reifying injustice [1]. This places a forward-looking responsibility not just on programmers, engineers, and manufacturers, but also on political actors, who have a responsibility to check in on and intervene in instances of algorithmically caused harm. This is also true of those who *fund* these programs, as funders also have an obligation to, for example, ensure that their projects adhere to trustworthy practices [9].

6 Beyond Consequentialist Reasoning

The second feature that might be thought to create in moral obligation is the *experimental* and *innovative* nature of many aspects of AI research. It is often the case, especially with machine learning systems, that the correlations they generate are *novel*. Thus, in the process of training these systems, engineers and programmers cannot predict the kinds of results that will be generated. The process of creating and deploying these systems is often an iterative one, with tweaks being made here and there to avoid failure or undesirable outputs. There is therefore a sense in which those involved are *experimenting* with the available data in order to try and derive a meaningful pattern that could be put to work. This kind of work is also *innovative* in the sense that it often involves a unique commercial application of a process or product. We might take innovation to mean the "*commercialization* of *technological* inventions" [3]. Of course, innovation in a broader sense is also possible (outside of technological innovation), but for my purposes, it is enough that we come to see the novel use of AI-based systems as a species of technological innovation. With this in mind, I will aim to show that the innovative and experimental nature of AI does not create a gap in moral obligation. Although it might seem as though these features would make it impossible to predict the consequences of AI-based interventions, this does not exhaust our ability to assess technology and thus does not come to complicate our active moral obligations.

In a recent article, Armin Grunwald [10] argues that a wholly consequentialist method of Technological Assessment (TA) does not work. He insightfully suggests that we may evaluate nascent technologies not only on their potential consequences but also by looking at the *hermeneutic knowledge* that is available to us when performing such evaluations. Here we find support for the idea that technological assessment is more than merely consequentialist, and should also include a hermeneutic component, which looks at the *societal meaning* of the system [10]. This raises the possibility that concerns strictly grounded in the control condition, with respect to moral obligation and AI, might not undermine our responsibility practices. Specifically, it suggests that the experimental and innovative nature of AI research does not foreclose discussions of moral obligation for future states of affairs.

One of the key complicating factors when reflecting on the future obligations that the creators of technology might have comes from the so-called Collingridge Dilemma [8]. Basically, the dilemma is that when a given technology is still in the nascent stages of development, it is possible to influence the way it will develop significantly however, we lack knowledge of how the technology will affect society. Once the technology becomes 'embedded' in society, and we come to know its implications, however, we are then in a position where we are unable to influence its development. In essence, when change is at its easiest, the need for it cannot be foreseen, and when change is required, it is difficult to implement [10]. This dilemma, when applied to AI, is especially pernicious. This is because, with AI, we are not only dealing with new technology subject to the constraints of the Collingridge Dilemma, but also a kind of technology that is capable of *learning* from its experiences, and thus, in a way, lying beyond the scope of the Collingridge Dilemma. Such technologies, from their very beginning, might not offer us the safety and security of *ever* being able to significantly influence their trajectories once they are deployed. I will argue, however, that reflection on the potential implications of

technology is more than a reflection on its potential consequences. This means that even if we cannot fully anticipate the consequences of a given technology, it does not follow that we are foreclosed from having an active moral obligation towards it.

An interesting point of departure on this journey is to reflect on the German translation of technological assessment: *Technikfolgenabschätzung*. Within this word, we find *Folgen*, which, literally, translates to "consequences" in English. Here we see how embedded the role of consequences are in TA, and rightly so. Prospective knowledge (knowledge about the future) is by its very nature uncertain, and so we need to devise a means with which we can reduce this uncertainty. This leads to attempts to develop mechanisms to *anticipate* what the future might hold, which acts as a guide to how we might structure our decision making in the present, with respect to novel technology [10].

When assessing technology, however, what exactly are we evaluating? Grunwald offers us two potential answers, and then gives reasons to reject both. He first claims that perhaps TA is an assessment of *technology*. However, this construal misses the fact there is no such thing as technology *as such*, and that technology is always embedded in a given *social environment* [10]. Second, he suggests that TA might concern itself with the *consequences of technology*. "Predicting, estimating and foresighting the possible consequences of technology belongs to TA's core business" [10]. Grunwald provides two reasons for rejecting this claim. The first is that the consequences of technology are not *just* the consequences of technology: these consequences are the result of varied and evolutionary interactions between technical, social, and institutional factors. Second, the consequences of technology do not yet exist. Therefore, strictly speaking, TA cannot be about these consequences *per se*, but only about the expectations, projections, or imaginations of *what they might be* [10]. In this way, we come to see that when evaluating technology, it is not enough to simply state that we should be concerned with the *consequences* of a specific technology, but rather we must interrogate the "*imaginations of future socio-technical configurations*" [10]. What might these "imaginations" be? Well, Grunwald argues that they need to fulfil two conditions to be proper objects of TA:

(1) Involve relations with science and technology

(2) Demonstrate that the technologies under consideration possibly have societal meaning and significance [10]

These two conditions seem natural enough. The first condition is straightforward, and once again points out that we must be conscious of the link between science and technology. The second condition introduces "societal meaning and significance", which is incredibly important to understand if we are to have a coherent means to evaluate novel technology. This criterion takes us beyond the mere consequences of the technology and prompts us to ask questions regarding the society-wide effects we may come to observe. These technologies are not simply additions to pre-given social systems, but come to influence ethical, economic, and social aspects of reality [11]. That is, we are now tasked with excavating what the technology might *mean* and are thus engaging in a distinctly *hermeneutical* project.

This draws our attention to how the projections and visions of new technologies come to shape their development. In order for scientists and researchers to secure public funds, they must convince those in charge of those funds (who are often not experts in the field)

that their research will be of great importance. This often has less to do with the science or technology itself, but what these will *make possible*. In addition to this, breakthroughs in science and technology can themselves fuel the creation of new forms of meaning. For example, ultrasound scans made it possible for people to "see" the developing foetus in the womb for the first time [21]. However, the associated meaning of this technology is very different to the consequences of the technology itself. The consequences of the technology are that it makes visible what was once invisible. In addition to this, however, ultrasound also made it possible to determine the thickness of the nape of the neck of a foetus [22]. With this knowledge, parents and doctors can make an assessment of the risk that the child will be born with Down's syndrome. Thus, the technology provides a new framework for how we understand the foetus: no longer an invisible entity, but a medical subject understood in terms of disease. Significantly, such an understanding also brings to light our ability to *prevent* certain foetuses from being born, should we be able to diagnose any "abnormalities" early enough. Thus, ultrasound technologies come to mediate certain moral questions regarding abortion. This mediation role of technology is well documented and provides support to the hermeneutic approach outlined above. Instead of *only* looking at the potential consequences of the technology, we need to focus our attention on trying to give an adequate account of how we understand it. This understanding is never "stable", as it is an iterative process (often called the "hermeneutic circle"): once we take the time to understand the social meaning of a technology we do not come back to our original starting position. Rather, the process of uncovering meaning itself creates a kind of spiral, whereby new inputs are interpreted by society in a number of ways and come to influence our understanding of the technology in question.

One way to deal with unpredictability might be to focus on the decision-making procedure itself. Even if we have some unpredictability in terms of *outcomes*, we might nonetheless be able to mitigate such risks by focusing on the *process* of AI development. Such a switch in perspective has been proposed by Stilgoe *et al*. [19], where they propose their AIRR (Anticipation, Reflexivity, Inclusion, and Responsiveness) framework. The AIRR framework provides a mechanism by which engineers can be educated about their professional roles *and* the means of fulfilling their moral obligations in these roles. The focus of this framework is on the *process* of responsible research, and so the unpredictability of the system does not *necessarily* undermine the feasibility of such an approach.

What does any of this have to do with AI and moral obligation? The point of this section has been to show that when assessing a technology, unreliable knowledge about the consequences of that technology do not foreclose our ability to investigate the societal meaning that the technology may hold. Therefore, while it may be true that the creators of AI systems might not be able to fully appreciate what the consequences of their systems might be, they can still take the time to investigate their societal significance. For example, 'predictive policing' algorithms have been touted as a mechanism to assist law enforcement with determining their inspection priorities [23]. These systems are meant to increase the efficiency and efficacy of law enforcement processes by targeting 'high risk' areas and deploying more resources to those areas.

From a purely consequentialist perspective, we could say that the consequences of this technology would be greater policing in high-risk zones. However, hermeneutically,

we might start to ask whether this kind of system might reinforce or create new forms of discrimination. What will the effects be of increasing police presence in high-risk areas, when those areas are historically disadvantaged? Unfortunately, with the benefit of hindsight, we can see that the results have been damaging for those communities. Hyper-surveillance partly produces increased rates of recorded crime (in the form of more arrests for petty crimes, for example), especially when the police know that they are being deployed in areas and are on the lookout for criminal behaviour, creating a guilty until proven innocent scenario [2]. This might reinforce existing racial prejudice on the part of police officers (if they are serving in historically disadvantaged communities) and may increase resentment in the community among those who feel that they are being unfairly targeted. The point of this example is that before deploying such systems, we should not merely look at the consequences of the *technology*, we also have to critically investigate how the technology will be embedded and what that might *mean* to the communities whom it will affect. Additionally, such "predictive systems" themselves operate under the assumption that attempting to predict the future will not influence the present. As the example should illustrate, and as a hermeneutical perspective illuminates, this is simply false, "as the very practice of forecasting the future partly acts directly upon the world - machine prediction plays a part in creating what exists whenever such predictions inform decision-making" [2].

Thus, those who design and implement such systems can undertake such hermeneutical analyses. There is no gap in moral obligation due to AI because here we have a mechanism that can overcome the cause of the potential gap (in the form of unpredictable consequences). In this respect, designers and developers ought to regularly check that the AI in question is performing its task in a way that is aligned with various socially desirable values (respect for human rights, equality, sustainability, etc.). This would involve understanding the specific context in which the AI is embedded, as well as how the agents interacting with it understand it, and how it affects the communities and groups within its range of influence.

However, this is not to say that this would be easy, or even that isolated engineers would be able to fulfill these obligations without education and input from researchers in the social sciences. My proposal therefore requires inter- and trans-disciplinary work so that the given societal meaning of the system can be uncovered. Such a process *demands* a diverse and pluralistic approach to technological assessment. Additionally, it might seem excessively onerous that programmers or engineers have to undertake such a hermeneutic analysis. This is especially concerning if we reflect on the gap between theory and practice that is operative in the AI ethics debate at present [14]. However, it is beyond the scope of the present paper to go into much detail with respect to how we might go about implementing such a hermeneutical perspective in applied contexts. My point here has merely been to suggest a theoretical perspective which might, when applied, yield more governable form(s) of AI systems.

7 Conclusion

In this paper, I have argued that AI systems do not create a unique gap in forward-looking responsibility. I supported this conclusion by focusing on the nature of risks

when developing technology, and by showing that technological assessment is not only about the consequences that technology might have. By broadening the horizons of what constitutes technological assessment, with specific reference to societal meaning, I aimed to show that we can avoid a gap in forward looking moral responsibility. This does not mean, however, that forward-looking responsibility is not an *issue* when it comes to developing and deploying AI systems. In fact, given what I have said here, it should be clear that AI does indeed *complicate* our responsibility ascriptions. However, such complications do not lead to an insurmountable gap.

References

1. Birhane, A.: Algorithmic injustice: a relational ethics approach. Patterns **2**(1), 1–9 (2021a). https://doi.org/10.1016/j.patter.2021.100205
2. Birhane, A.: The impossibility of automating ambiguity. Artif. Life **27**, 1–18 (2021b)
3. Blok, V.: 'What Is Innovation? Laying the ground for a philosophy of innovation. Techné Res. Philos. Technol. **1**, 1–25 (2020). https://doi.org/10.5840/techne2020109129
4. Boenink, M., Kudina, O.: Values in responsible research and innovation: from entities to practices. J. Responsible Innov. **7**(3), 450–470 (2020). https://doi.org/10.1080/23299460.2020.1806451
5. Bryson, J.J.: Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics Inf. Technol. **20**(1), 15–26 (2018). https://doi.org/10.1007/s10676-018-9448-6
6. Bryson, J.J.: The artificial intelligence of the ethics of artificial intelligence: an introductory overview for law and regulation. In: Dubber, M., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, New York (2020)
7. Carrier, M.: How to conceive of science for the benefit of society: prospects of responsible research and innovation. Synthese **198**(19), 4749–4768 (2019). https://doi.org/10.1007/s11229-019-02254-1
8. Collingridge, D.: The Social Control of Technology. Frances Pinter Limited, London (1980). https://doi.org/10.2307/1960465
9. Gardner, A., Smith, A.L., Steventon, A., Coughlan, E., Oldfield, M.: Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. AI Ethics **15**, 1–15 (2021). https://doi.org/10.1007/s43681-021-00069-w
10. Grunwald, A.: The objects of technology assessment. Hermeneutic extension of consequentialist reasoning. J. Responsible Innov. **7**(1), 96–112 (2020). https://doi.org/10.1080/23299460.2019.1647086
11. Henry, N., Powell, A.: Sexual Violence in the Digital Age, Social and Legal Studies. Palgrave Macmillan, London (2017). https://doi.org/10.1177/0964663915624273
12. 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
13. Mecacci, G., Santoni de Sio, F.: Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. Ethics Inf. Technol. **22**(2), 103–115 (2019). https://doi.org/10.1007/s10676-019-09519-w
14. Morley, J., et al.: Ethics as a service: a pragmatic operationalisation of AI ethics. Minds Mach. **31**, 239–256 (2021). https://doi.org/10.2139/ssrn.3784238
15. Nyholm, S.: Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. Sci. Eng. Ethics **24**(4), 1201–1219 (2017). https://doi.org/10.1007/s11948-017-9943-x

16. van de Poel, I., Sand, M.: Varieties of responsibility: two problems of responsible innovation. Synthese **198**(19), 4769–4787 (2018). https://doi.org/10.1007/s11229-018-01951-7
17. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos. Technol. **34**, 1057–1084 (2021). https://doi.org/10.1007/s13347-021-00450-x
18. Shapin, S.: The Scientific Life: A Moral History of a Late Modern Vocation. The University of Chicago Press, Chicago (2008). https://doi.org/10.1002/sce.20372
19. Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation. Res. Policy **42**, 1568–1580 (2013). https://doi.org/10.1002/9781118551424.ch2
20. Tigard, D.W.: There is no techno-responsibility gap. Philos. Technol. **34**(3), 589–607 (2020). https://doi.org/10.1007/s13347-020-00414-7
21. Verbeek, P.P.: What Things Do. The Pennsylvania State University Press, University Park (2005). https://doi.org/10.1017/CBO9781107415324.004
22. Verbeek, P.P.: Materializing morality: design ethics and technological mediation. Sci. Technol. Human Values **31**(3), 361–380 (2006). https://doi.org/10.1097/EDE.0b013e3181
23. Yeung, K.: "Hypernudge": big data as a mode of regulation by design. Inf. Commun. Soc. **20**(1), 118–136 (2017). https://doi.org/10.1080/1369118X.2016.1186713