AI & SOCIETY
https://doi.org/10.1007/s00146-022-01498-9

ORIGINAL PAPER

Expert responsibility in AI development

Maria Hedlund[1] · **Erik Persson**[2]

Received: 20 August 2021 / Accepted: 4 April 2022

Abstract
The purpose of this paper is to discuss the responsibility of AI experts for guiding the development of AI in a desirable direction. More specifically, the aim is to answer the following research question: To what extent are AI experts responsible in a forward-looking way for effects of AI technology that go beyond the immediate concerns of the programmer or designer? AI experts, in this paper conceptualised as experts regarding the technological aspects of AI, have knowledge and control of AI technology that non-experts do not have. Drawing on responsibility theory, theories of the policy process, and critical algorithm studies, we discuss to what extent this capacity, and the positions that these experts have to influence the AI development, make AI experts responsible in a forward-looking sense for consequences of the use of AI technology. We conclude that, as a professional collective, AI experts, to some extent, are responsible in a forward-looking sense for consequences of use of AI technology that they could foresee, but with the risk of increased influence of AI experts at the expense of other actors. It is crucial that a diversity of actors is included in democratic processes on the future development of AI, but for this to be meaningful, AI experts need to take responsibility for how the AI technology they develop affects public deliberation.

Keywords Artificial intelligence · AI · Responsibility · Forward-looking responsibility · Experts · Democracy

1 Introduction

The aim of this paper is to discuss the extent of responsibility of technological expertise for guiding AI development in a desirable direction. Intuitively, it makes sense to claim that AI experts (in this case, experts of the technology side of AI) are responsible for the technology they develop, but the question is to what extent they are responsible, and for what. In this paper, we elaborate on these issues by trying to answer the following research question: *To what extent are AI experts responsible in a forward-looking way for the direction of AI development that go beyond the immediate concerns of the programmer or designer?*

Artificial Intelligence (AI) has its roots in the 1950s, but it is only recently that AI development has become a veritable boom and AI has become a buzzword that actors from practically all sectors in society refer to. No one wants to be left behind. Corporations and decision-makers all over the world strive for leading positions, or at least not to miss the boat. Under these circumstances, AI experts are hard currency, frequently asked for as well by the industry to develop AI, as by decision-makers to act as advisers in ethical committees. Their technological knowledge makes AI experts better than the rest of us at everything from the potentials of AI technology to how the details in the codes should be written to achieve a certain effect. All this puts AI experts in a position to have a significant influence on the direction of AI development.

The AI technology that these experts develop often comes with huge benefits for society. Self-driving vehicles can make transportation of people and goods safe and effective, and machine learning technology enables image recognition for accurate and effective medical diagnosis, search engine recommendations, or automated financial trading to mention just a few examples of how AI development has made our daily lives more comfortable. Other AI applications can increase efficiency of farming, contribute to climate change mitigation and adaptation, or improve efficiency in production.

✉ Maria Hedlund
maria.hedlund@svet.lu.se

Erik Persson
erik.persson@fil.lu.se

[1] Department of Political Science, Lund University, Lund, Sweden

[2] Department of Philosophy, Lund University, Lund, Sweden

Published online: 13 June 2022

However, AI development also entails risks. Some of these risks are connected to the special characteristics of AI compared to other technologies: its capacities to self-improve and to act autonomously. Surveillance systems aimed at decreasing crime and increasing security sometimes violate privacy and individual freedom; algorithmic bias sometimes leads to discrimination; algorithmic content recommendations may increase polarisation and risk harming democratic deliberation; the dominance of a few AI companies, large enough to generate their own massive datasets, not only makes it difficult for smaller competitors to match, but also gathers huge amounts of data about individuals all over the world in a few hands, with the risk of misuse that follows with monopolising orientation; and authoritarian tendencies around the world seem to be accelerating by the help of AI systems.

Although AI experts typically do not intend to cause these harmful effects—as in all areas, some AI experts might have bad intentions, but our focus here is on the serious ones—the question is if they could foresee the negative effects, or, rather, if they should foresee them. Do AI experts have a responsibility to anticipate possible unintended consequences of the technology they develop, such as misuse, or, even more intricate, side effects of the very functioning of the technology when it works as intended?

We will start by explaining the key concepts of 'AI experts', 'desirable direction', and 'forward-looking responsibility', and then elaborate on the extension of expert responsibility within and beyond the policy process. As we will see, discussing responsibility for AI experts in terms of individual responsibility is insufficient. We will also explain, and, to some extent, elucidate two of the core problems, namely collective responsibility and the problem of many hands. To illustrate the complexity of attributing forward-looking responsibility for unintended, undesirable consequences, we use the examples of recommender algorithms.

2 Who is an AI expert?

An expert is commonly referred to as someone who possesses specialist knowledge that is accepted by a wider society as legitimate (Schudson 2006; Goldman 2006; Hoard 2015; Watson 2019). Such a broad definition allows for many different kinds of actors to be covered by the notion of AI experts: researchers and professionals working with AI development as well as researchers studying AI development, policy-makers, and other actors involved in policy processes on AI development. In this paper, focus is on the responsibility of actors with first-hand knowledge of AI technology—the handicraft—of the development of AI technology. More specifically, in this paper, the term 'AI experts' refers to AI researchers (e.g. computer scientists, mathematicians, and cognitive scientists), developers, designers, programmers, engineers and other technology professionals working with AI. These experts play important roles in the development of AI technology and the functioning of applications built on AI technology, but also in policy-making on AI development in several settings: in corporations and professional networks; at universities and research institutes; and as advisors in political processes. In other words, AI experts are well in a position to influence AI development.

3 What does it mean to lead AI development in a desirable direction?

Our forward-looking approach makes concern for future generations crucial (Scheffler 2018). While there are several candidate values that could characterise a desirable future—peace, justice and wellbeing of all may be some obvious such values—we argue that safety and democracy are critical values in this regard. Since we have a long-term perspective, we need to consider that prioritisation of values change over time (c.f., Kudina & Verbeek 2018). Future people may appreciate other values than people living today. To enable future generations to be able to decide on their own values, we suggest that the best thing we can do today is, first, to make it probable that there is a future for humans and other sentient beings, and second, to make it probable that also people in the future will be able to decide which values should be promoted. Thus, for our purposes here, we adopt an absolute safety concept (Hansson 2012), referring to the survival of sentient, biological life, including but not limited to human beings. Democracy, an important basis for people's ability to form and pursue their own values, will here refer to a system in which citizens have equal rights and real possibilities to decide on common matters (Dahl 1989). We believe that steering the development of AI in a direction that promotes safety and democracy enables every member of society also in the long run to decide which values to be promoted.

4 Forward-looking responsibility

When talking about moral responsibility,[1] it is common to mean retrospective or backward-looking responsibility, referring to a causal link between you and something that

[1] Our focus is on moral responsibility and not on legal responsibility, which gives rise to partly different considerations. If nothing else is explicitly pointed out, responsibility in this text refers to moral responsibility.

has already happened, and usually also for being blameworthy (or praiseworthy) for having acted (or not having acted) in a certain way. Common requirements for backward-looking responsibility are that the actor has freely and knowingly caused the blameworthy (or praiseworthy) situation. In this paper, focussing on the future development of AI, the focus is on prospective or *forward-looking* responsibility. While backward-looking responsibility is about things that have happened in the past, forward-looking responsibility is about obligations regarding something that is not yet the case, or the duty to act (or not to act) in a certain way (van de Poel 2015a: 27; Knaggård et al. 2020; Persson et al. 2021).

Requirements for being responsible in a forward-looking way differ from requirements for being responsible in a backward-looking sense. One thing is that causality plays a somewhat different role. An actor could be responsible in a forward-looking way to remedy or mitigate a harm on the basis that she has previously caused this harm, but in the forward-looking respect, causality need not play a decisive role (e.g. Young 2006, 2011). Understanding forward-looking responsibility as an obligation to bring about a certain outcome—in our case a democratic and safe society—does not necessarily mean that the responsible actor herself brings this outcome about, only that she is responsible for this outcome to happen (van de Poel 2015a: 29). This looser causality requirement indicates that there are other possible grounds for distributing forward-looking responsibility.

Depending on ethical theory, forward-looking responsibility could be distributed according to different criteria. As we are interested in a certain outcome—a democratic and safe society—consequentialist criteria are appropriate. In this context, dealing with expert responsibility, we believe that *capacity* and *position* are useful bases for the assignment of forward-looking responsibility. In addition, we believe that *influence* is a necessary requirement for responsibility; it is not possible to be responsible for something you cannot influence (Kant 1999). The principle of capacity holds that forward-looking responsibility should be assigned according to the capacity of each agent to affect the future situation in a desirable way (Miller 2001: 460). Capacity so understood could be seen as an ability or a competence. Position refers to being at the best place to bring about a certain outcome (Miller 2001: 454), for instance by being a bystander observing someone falling over and getting hurt, or being involved in policy-making processes. Position, so understood, may, but need not, overlap with capacity.

5 Expert responsibility

Generally, the responsibility of experts does not differ from the responsibility that all of us have to bring about a desirable future (Douglas 2009: 71). However, in relation to non-experts, experts have superior capacity within their expert domain, implying that they have a larger responsibility to bring about a certain outcome within this domain. The context also matters. In political decision-making processes, experts are called on to provide their expertise, while in their ordinary work as scientists, AI developers, or programmers, they expand and/or make practical use of their expertise in a way that has direct or indirect societal consequences. Contrary to what is sometimes claimed (e.g. Walsh 2018: 11), technology is not value-neutral (Fossa 2018: 122; Dignum 2019: 94). Like all technologies, AI technology shapes our actions and contributes to our moral decisions (Verbeek 2006: 366; Schraube 2009). This suggests a special responsibility for those who design the technologies (Verbeek 2006: 369), in our case the AI experts.

Thus, AI experts have the capacity to influence AI development and many of them are in a position to do so, within or beyond policy processes. Within the policy process, AI experts are called on to provide knowledge and to act as policy advisors. One current example is the High-Level Expert Group on Artificial Intelligence, a constellation of 52 AI experts from industry, academia, and civil society.[2] On commission by the European Commission, the group provided ethical guidelines and investment recommendations, advice that will constitute a basis for EU regulation on AI (AI HLEG 2019a, 2019b). Beyond the policy process, AI experts directly and indirectly influence the technology as such. In addition, they exert influence on how people think about AI technology through scientific and popular publications; by organising conferences and exhibitions; and by talking to people at the workplace, in the family, in social media and in other contexts where they, as all of us, meet other people. However, our interest here is primarily the influence of AI experts in policy processes and in their work as developers, designers, programmers and so on. What does this influence imply for their forward-looking responsibility? How far beyond their immediate action does the forward-looking responsibility extend?

AI experts formally involved in policy processes have a real—but not always formal—possibility to influence regulation and legislation (Jasanoff 1990; Lindvall 2009). If we accept influence as a basis for responsibility, AI experts involved in policy processes have a responsibility to steer the direction of AI development in a desirable direction. Policy processes on AI are definitely important for the future development of AI, but as is pointed out above, AI experts have substantial influence when they are doing hands-on work with AI development by programming, designing,

[2] Although all these experts were not technological experts in the meaning we use in this article, a majority of them were (EG Regexpert; EC Futurium; Metzinger 2019).

engineering and so forth. Such activities beyond the policy process also contribute to steering the direction of AI development. Coding and design produce certain functions that affect our behaviour and the way societal institutions work, or, in Verbeek's words: engineers and designers are doing "ethics by other means" (2006: 369). This is a reason why it is important also to consider effects beyond the immediate concern of the AI expert.

Experts who are formally involved in policy processes have different options: they can present facts or possible lines of action, or they can make recommendations (Gundersen 2018). In neither case are they decision-makers. This means that they are exempt from formal (backward-looking) responsibility, but morally, they have forward-looking responsibility. Experts are always responsible for the quality of their contributions (Matzner & Barben 2020: 75–76), and it is reasonable that they are morally responsible for what they recommend. Moreover, though they cannot be responsible for what the decision-makers choose to do with their recommendations, experts consulted to provide specialist knowledge in policy processes often play a de facto political role (Metz 2015; Van Ballaert 2015).[3] Recommendations are per definition normative. It can even be claimed that also presenting facts or possible lines of action is a normative endeavour—there is always a choice of which facts to present. This latter aspect is central in Douglas' (2009) theory of the responsibility of scientists.

Douglas (2009) asserts that, in addition to the responsibilities that come with being a scientist, such as adhering to basic principles of research ethics, scientists are also responsible for the consequences of their empirical claims. More specifically, she argues that scientists are responsible in a forward-looking sense for the foreseeable consequences of their claims, whether intended or not (2009: 71). This applies in one particular situation, namely, when they consider the consequences of error in their work. According to Douglas, scientists are responsible to balance potential unintended consequences of making inaccurate or unreliable empirical claims. As "there can be clear consequences for making a well-intended but still ultimately incorrect claim" (2009: 72), scientists should balance the effects of possible outcomes before they advise decision-makers; the usefulness of a general requirement of wearing a face mask or not during a pandemic may be a case in point. Key for Douglas' argument is how scientists should deal with uncertainty when they make empirical claims. What consequences are preferable? Claiming that masks prevent the spreading of virus if they turn out not to prevent, or even worsen, the spread of the virus? Or claiming that masks do not prevent the spreading of viruses if they turn out to actually do so? Uncertainty about the protection is the key here. In both cases, the error would have some bad effects on the spread of the virus, and for Douglas, it is the responsibility of the scientist to weigh such consequences before they make the claim:

> Less uncertainty is tolerable when the consequences of error are [...] high, and in making the claim that the risk of catastrophe is low, we would want a high degree of surety. This *is* the weighing of uncertainty around an empirical claim against the ethical consequences of error (Douglas 2009: 78).

Following Douglas, scientists are the ones morally responsible to balance the potential consequences of an empirical claim that may be incorrect. Someone has to make the choice, and compared to non-scientists, scientists have the best capacity to make this choice, and thus, they have this responsibility. In addition, it could be argued that scientists have this responsibility by virtue of their role or position (Young 2006: 119).[4]

For our purposes here, two questions arise. First, if we accept that scientists have a responsibility to balance potential consequences of error before they make an empirical claim, could we then extend this responsibility also to experts who are not scientists? We believe that to be a reasonable extension. Arguments for assigning forward-looking responsibility to scientists regarding potential consequences of errors in scientific work are their superior capacity in this regard, and the authority that science, because of this superiority, enjoys in society (Douglas 2009: 82). Similar arguments could be put forward regarding AI experts who are not scientists. Their knowledge about AI technology—how it works, prospects for further development, possible applications—is very specialised and not easily attainable for people outside the AI domain. Thus, also AI experts who are not scientists possess a superiority within their area of specialisation that is comparable to the superiority of scientists in their areas of expertise, and it is not unreasonable that AI experts have responsibility analogous to that of scientists. Second, how about unintended consequences that do not depend on potential error, but, on the contrary, occur just because things do work as intended?

[3] However, if this political role is an effect of how the experts are used by their principals, it is not obvious that it affects the responsibility of the experts. It might simply prove that the experts are not only used for providing specialist knowledge (c.f., Boswell 2009). On the other hand, as it opens up the possibility for experts to have a real influence, it is not unreasonable to assert expert responsibility on this basis.

[4] This standpoint may seem self-evident, but the claim that scientists should be responsible for consequences of the knowledge they produce is contested. Many philosophers of science have assumed that scientists should be insulated from moral considerations (for an overview, see Douglas 2009: 44–65).

6 Unintended consequences

Unintended consequences refer to results undertaken for another purpose, irrespective of whether they are evaluated as good or bad (Lidskog and Sjödin 2018: 3). As our focus is on potential harm resulting from AI technology that works as intended, we qualify the question to apply to unintended consequences that are undesirable. Thus, we pay attention to undesirable consequences that are unintended. Although there surely exist AI experts with bad intentions, from a deontological point of view, their actions may be directly blameworthy no matter if they lead to undesirable consequences or not. However, that is beyond our scope here. More intriguing, and in focus of this paper, are bad effects resulting from something that is done with *good intentions*. To what extent should AI experts be responsible for preventing unintended, undesirable consequences of AI technology that functions exactly as it is supposed to do? Consider for example search algorithms that adapt to individual search patterns. By learning from previous user behaviour, the search algorithm produces results that align with what the user has searched before. This certainly makes searching more effective, but at the same time, the search result is not only narrowed down, but also skewed in some way. This bias may not only affect the individual user, but also have effects on a societal level. Consider a search for the term 'vaccination'. Due to previous searches, an individual sceptical towards vaccination may get search results that discourage vaccination, which is completely pertinent to this user, but may be counterproductive on a societal level (Haider and Sundin 2020: 10). The algorithm has worked as it is supposed to do, but has undesirable side effects. To what extent should AI experts be responsible for avoiding such effects?

In general, responsibility theory points at two aspects of importance for deciding whether an actor is responsible for preventing unintended consequences of their actions: recklessness and negligence (Feinberg as referred to by Douglas 2009: 68–69). Someone who acts *recklessly* is aware of risks that a particular action imposes on others, but acts anyway. However, risks could be deemed to be justified, for instance if you exceed the speed limit to take an injured person to a hospital. But if you exceed the speed limit just for the fun of it, your risky behaviour most probably is deemed to be unjustified. An actor who is aware of the unjustified risk of an action, and still goes ahead, is irresponsible due to recklessness. Someone who acts *negligently*, on the other hand, is unaware of the risks, but should have been aware. This means that a reasonable person reflects on the risks, while a negligent person does not bother to think about potential consequences of their action. We suggest that an expert has a responsibility to avoid both unjustified recklessness and negligence that may lead to unintended consequences; the difference lies in the agent's thought processes. To some extent, we are all morally responsible to avoid harmful side effects of our actions. The crucial question here is to what extent AI experts are responsible in a forward-looking sense for avoiding potential harmful side effects of the technology they develop, and whether they, due to their superior knowledge and skills in this area, should carry a heavier responsibility burden than the rest of us for these effects. We have already argued that experts have a special responsibility within their area of expertise by virtue of their expertise, which equips them with superior capacity to bring about a certain outcome. Hence, they have a larger responsibility than the rest of us to avoid potential harmful side effects of the technology they develop, given that they could foresee them.

Responsibility to avoid unintended consequences gives rise to methodological problems regarding causality and intention. What consequences can justifiably be attributed to a certain situation? In addition, how could we ascertain the actual purpose of a given situation (Lidskog and Sjödin 2018: 3)? Causal connections certainly play a role for forward-looking responsibility, especially as we regard influence as a basis for responsibility, but causality—or efficacy—alone is not sufficient. For our objective, it suffices to state that AI experts by their actions give rise to AI technology that have certain effects, directly or indirectly, intended or unintended. However, as we will discuss further below, AI experts do not act in a vacuum. Investors and consumers certainly contribute to the direction of AI development by demanding certain functions. Nevertheless, without the knowledge and expertise of AI experts, these functions would never materialise. An example will help illustrate the complexity of attributing forward-looking responsibility to avoid unintended and undesirable side effects.

7 Recommender algorithms and impaired democratic debate

Consider the algorithmic selection bias, so-called 'filter bubbles' or 'echo chambers', that emerge as an effect of how algorithms recommend users to content that is supposed to maximise platform usage (Sirbu et al. 2019). While the rationale behind such algorithms is to sell targeted advertisements, the effects may go far beyond the displaying of ads that attract the users' attention and, eventually, lead to purchase. Based on what users 'like' or share on social media platforms, or on what they search in search engines (see the example above about vaccination), these algorithms direct them to content that they are assumed to prefer. With

a market share between 85 and 90% (Haider & Sundin 2020: 11), Google search dominates the market of search engines.[5] By this dominance, its search results play a considerable role in shaping the perception of the world. Since search results are adapted to user behaviour, and since users get suggestions for further, related searches, this perception does not necessarily reflect a representative picture of the world, but a picture in which "the known becomes more known" (Haider and Sundin 2020: 5) and which differs between users. As with the recommender algorithms in social media, the search algorithms have worked as intended—effectively providing the user with pertinent search results—but also in this case, the effect is that we are confronted with different world views, which makes democratic debate and mutual tolerance more difficult.

In that way, users are likely to see more of what they are already seeing, and may not even encounter narratives different from their favourite one. This means that users do not get confronted with much resistance to the views they hold and may get the impression that everybody agrees with them. This is detrimental for democratic deliberation in several respects: exchange of different opinions, which democratic discussion basically is about, disappears; tolerance for diverging views, a basic requirement for a truly democratic society, may decrease; the idea that deliberation—in the ideal, Habermasian sense—refines standpoints is even more difficult to implement; and John Stuart Mill's idea of deliberation as a way to sort out faults and find the truth would be impossible. The algorithm has worked as intended, but when implemented in real-life situations, it has unwelcome consequences. Is the programmer of the recommender algorithm that has this detrimental effect on democracy responsible for foreseeing and preventing this effect?

This question points to two separate but related aspects of importance for responsibility assignment. First, as we notice above, technology is never neutral, but incorporates judgments (Jasanoff 2016: 11; Schraube 2009) and "play an active role in the relationship between humans and their world" (Verbeek 2006: 365). Designers and engineers make decisions about characteristics of the technology they develop, decisions that have consequences for users (Verbeek 2006: 362). Moreover, it could be argued that we should not strive for technology to be neutral, but be designed "in a way at least consistent with the public health, safety, and welfare" (Davis 2012: 31). Like the speed bump is designed to make the driver decrease the speed, the social media algorithms are programmed to lead the user to content they have already shown preference for. The echo chamber effect that this entails may have consequences that go far beyond the immediate comfort that makes the user stay longer and make more clicks. For Verbeek, this implies that "technologies can contribute to the moral decisions human beings make" (2006: 366) and creates "a specific responsibility for designers" (2006: 369). Following this way of arguing, our AI expert is responsible for what the algorithm makes people do, but it does not tell us how far beyond the immediate action this responsibility extends.

This leads us to the second aspect of importance for responsibility assignment that the recommender algorithm case points at, namely unintended consequences. It is rather straightforward to assert that the programmer is responsible for the direct effect of the algorithm; the programmer clearly is in the position to make a certain outcome more plausible. Moreover, the focus on unintended side effects makes this direct causal chain only a first step: the intended effect to make the user stay longer. An unintended consequence of this effect is the generation of the 'echo chamber', which affects the process of opinion formation in detrimental ways: by inducing fragmentation and clustering of opinions, and by intensifying polarisation and increasing distance between opinions (Sirbu et al. 2019: 2).[6] Unintended side effects are often referred to when errors occur, either due to bad execution—the original intent was not executed as planned—or due to things happening outside the scope of the original intent (Jasanoff 2016: 24). In this case, no error occurred; the recommender algorithm worked as intended. Rather, things happened outside the scope of the original intent, and the question is whether our AI expert should have foreseen this consequence of the algorithm and have a responsibility to minimise the harms of the democratic debate that the algorithm indirectly gave rise to. While the responsibility of AI experts may be constrained by their commission, legislative regulations, and organisational norms (Orr and Davis 2020), their constructions have effects also on other things, such as democracy, and following our reasoning, it could be argued that they are responsible also for these effects.

For those of us who are not AI experts, it is difficult to judge whether such a consequence would have been foreseeable, which points to the novice/expert problem (Goldman 2006: 18), referring to the difficulty or impossibility for a non-expert to evaluate the expertise of experts. This suggests that we have no choice but to rely on experts (Bohman 1999; Baier 1986). Hence, whether the AI expert claims that she could have foreseen this undesirable, unintended consequence, or that she could not have foreseen it, it seems that we have no choice but to trust her. In line with our argument so far, in the former case, the AI expert is responsible, in the

[5] In the Western world. Russia and China have their own search tools (Haider & Sundin 2020: 11).

[6] Although the viability of democracy benefits from the existence of different opinions, for democratic debate to be feasible and meaningful, some common understanding is crucial (Manin et al. 1987).

latter case she is not. However, it is not completely satisfying that an actor can get away from responsibility simply by claiming lack of ability.

It can also be suspected that in some cases unintended side effects would be foreseeable if the AI experts were really set on discovering them. In these cases, the real explanation for why damaging side effects are not foreseen is not that they are unforeseeable, but that effects, other than those directly relating to the bottom line of the company, are seen and treated as externalities. This, in turn, means that those who would be best positioned to foresee them, the AI experts, are not encouraged to bother about externalities or with thinking of ethical and societal aspects of their work. We will look at such a case in a later section. Here we can conclude that the kind of responsibility we promote for AI experts, that is, forward-looking responsibility for contributing (or at least not disrupting) a safe and democratic future, implies that AI experts do have a responsibility to spend reasonable effort beyond the law and beyond the instructions from their managers or customers to foresee effects on safety and democracy of their constructions.

Although one single actor may not possess the knowledge and overview necessary to fully understand the consequences on the societal level of her actions, she could be responsible to get the knowledge and overview needed to be able to foresee the undesirable, unintended consequences on democracy of her work. But is this effect on democracy something that the programmer could do something about? Probably not in the immediate practice. But should she do it? On one hand, we have a societal interest in a well-functioning democracy. On the other hand, we have the interest of the employee to do her duty to her employer. (Which, as it is, also puts bread on the table.) We clearly have interests of different levels here: the balancing between short-term individual interest and long-term societal interest. Under what circumstances we should expect an individual to make personal sacrifices for the common good is a question that somehow applies to all of us as members of a society, and to investigate those circumstances is beyond the aim of this paper. Here, the issue is if there are any arguments for a special responsibility for AI experts when it comes to unintended side effects of the technology that they develop.

What we are trying to get hold of is whether the responsibility that AI experts have by virtue of their expertise gives them a special responsibility for the direction of AI development, but it is important to point out that we do not suggest that the AI expert is the only candidate for responsibility. On the contrary, a number of different actors in society play a part here. In addition to experts of the technological aspects of AI, that we are talking about here, also experts on AI from other perspectives than technology, for instance, policy-makers, interest organisations, media, educators, investors, users of AI technology, researchers that study ethical, social and other non-technology aspects of AI, and AI corporations that hire AI experts, are important actors in AI development. To some extent, the individual programmer is responsible for the formulation of the code, but she acts in a context. There may be many coders working together. They have a boss who tells them what code to write. They work for a company that has a certain idea of what to accomplish, in which this code is just a part of a bigger system, in which corporations strive to meet expectations from consumers and shareholders (c.f., Orr and Davis 2020). Certainly, there is a connection between the work of the individual programmer and the effects on the AI system that she is contributing to, but it is just a part of an entirety. This makes it difficult to determine the influence of this programmer. In addition, the individual programmer's contribution may not play a role at all if not considered within the context in which it is produced. Responsibility would need to be assigned collectively to all actors contributing to this AI system. But collective responsibility is a notoriously difficult concept, as being morally responsible requires moral agency, and it is not completely clear under which circumstances, if any, a collective qualifies as a moral agent.

8 Collective responsibility and the problem of many hands

Collective responsibility is a contested matter, as collective agents do not have the control over their actions that is crucial for moral responsibility, and in particular, since collectives are not sentient entities and can thus not have intentions or care about the outcome of their actions in the same sense as individual sentient beings. When the responsibility of a collective is distributive between the members of the collective, and when this distribution is morally justified, it could be claimed that we are faced with collective responsibility that is analysable in terms of individual responsibilities (Held 1970: 97; van de Poel 2015b: 65–66). However, collective responsibility that is analysable in terms of individual responsibilities seems most often to be discussed in connection to backward-looking responsibility, which is beyond the focus in this article. Moreover, it could be questioned whether we are really faced with *collective* responsibility in such cases.

Collective responsibility that is *non-distributive*, on the other hand, refers to situations when the responsibility of the collective is more than the sum of the individual responsibilities of its members, i.e. when only the group as a whole, by joining in collective action, could prevent the harm (Feinberg 1970; Muyskens 1982; Young 2006). While distributable responsibility rests on the moral agency of each of the individual members of the collective, non-distributable collective responsibility brings back the question of collective

moral agency: can a collective be responsible as a collective, without reducing the responsibility to individual responsibilities? This situation, also known as the problem of many hands, arises when it is "reasonable to hold a collective, as a whole, morally responsible" while "none of the individuals meets the conditions for individual responsibility" (van de Poel 2015b: 52). In such a situation, each of the individuals could well be individually responsible for a particular subset of the whole, but as the whole is constituted by these subsets taken together, or even more than that, no individual could be assigned responsibility for the whole.

When the problem of many hands applies, an alternative to collective responsibility may be that no one could be attributed responsibility, which, in turn, entails the risk that "no one needs to feel responsible beforehand, meaning that future harm cannot be prevented" (van de Poel 2015b: 52). Hence, there might be a good reason to accept the notion of collective responsibility. Nevertheless, the problem of moral agency remains: is it reasonable to regard a collective as a moral agent, and if so, under which conditions?[7] To be a moral agent, an actor must have awareness of the moral nature of an action. If a collective has such awareness, then the requirements to hold a collective morally responsible is the same as for holding individuals responsible (Held 1970: 90–91). While a collective cannot have awareness in the same sense as individuals, one could instead talk of collective aims:

> A collective aim is an aim in the minds of individuals, but its content may be irreducibly collective in the sense that it refers to things that can only be collectively achieved and not by individuals in isolation (van de Poel 2015b: 56).[8]

The problem of many hands basically implies that no one is responsible and typically arises when collectives cannot make a joint decision or do not have a collective aim. While not all collectives have aims in this respect, *organised groups* like states, companies, and universities, are characterised by institutional rules and decision procedures "by which certain actions or aims can be authoritatively represented as the actions or aims of the collective" (van de Poel 2015b: 57; c.f., Held 1970). *Joint actions*, such as travelling together, or a research project, are actions undertaken by a group of individuals jointly to achieve a collective aim. When group members intentionally contribute to the joint action by doing particular parts, responsibility could be assigned to that collective (van de Poel 2015b: 59). *Professions* is a kind of collective that makes the issue of responsibility less problematic than for other collectives (Hedlund 2012). A profession is chosen, and with this choice, one adopts the values, customs, and rules of the profession, and assumes the responsibilities connected to being a professional. Moreover, every profession holds exclusive collective power or rights, coupled with collective duties to maintain professional standards and ideals (Muyskens 1982: 172–173).

9 Are AI experts collectively responsible for avoiding unintended consequences?

Let us, tentatively, regard the AI experts as a professional collective (remember that we are here only dealing with a specific set of AI experts, viz. experts of AI technology). Would it, then, be reasonable to assign forward-looking responsibility to this collective? Returning to the social media recommender algorithm, the intended function is, as noted above, to maximise attention time to sell advertisements, not—one would hope—to cluster users into isolated bubbles without any contact with people with a different view. This is an unintended side-effect, with harmful effects on democracy. We have discussed the difficulty to assign forward-looking responsibility in this regard to an individual AI expert unless she is responsible to get the knowledge required to be able to foresee these side effects. But since the contribution of the individual programmer constitutes only one small part of a bigger whole—programmers often work in networks, or are hired on a freelance basis to provide a certain piece of code—it could be unreasonable to expect this overview and foresight from one single individual. Under such circumstances, it would make more sense to expect AI experts as a professional collective to foresee these consequences. As implied above, this question could be seen as a question of knowledge. Even when it is reasonable to expect the individual to get the knowledge required for responsibility, this knowledge would be limited compared to that of a collective. If we understand knowledge of a collective as "the knowledge that would become known if all group members shared their knowledge effectively with all other members" (van de Poel 2015b: 62), and if we consider that the number of experts working with AI—the profession of AI—is rapidly growing,[9] the scope of knowledge and oversight increases dramatically. It is probable that AI experts as a *collective* have the knowledge and oversight to be able to foresee potential harmful side effects of the

[7] This question is discussed among others by Erskine (2015), Lang Jr (2015), Miller (2001).

[8] Although this seems to point more at the outcome, it could be argued that individual actions aiming at the achievement of a collective end can be attributed collective intentions (Kutz 2000: 85–89).

[9] The hiring of AI specialists has grown 74% in the last four years (*Computer Science* 2021), there is a forecast of 58 million new jobs in AI by 2022, and according to Linkedin, AI specialist is the top emerging job in 2020 (*Forbes* January 5, 2020).

technology they develop, including detrimental effects on democracy, and if they do, in line with our reasoning above, they have responsibility in a forward-looking sense to avoid these side effects. However, this responsibility cannot mean that the profession of AI experts acts on its own accord. Let us illustrate with an example.

A frequently suggested solution that is highly dependent on the skills and knowledge of the experts is to make the technology itself ethical, what Dignum (2019: 7) refers to as "ethics by design" or, as Verbeek (2006: 369) puts it, to inscribe morality in the product. This direction, also indicated by the mentioned High-Level Expert Group in its ethical guidelines,[10] has been criticised for reducing human freedom (Verbeek 2006: 369). A possible objection to this criticism is that technologies always shape human actions. Designing them to shape the actions in a desirable way is preferable to shaping actions in a bad direction or without consideration of how they shape human action. Seen from this perspective, "ethics by design" could be seen as a way of taking forward-looking responsibility for the effects of AI development.

However, there are also other problems with this approach. One issue with a technical solution to avoiding harmful side effects is that it may increase the influence of AI experts (Persson and Hedlund 2021). A crucial question is which values should be promoted by the technology and how this should be settled. We certainly do not want programmers themselves to make their own decisions on which values to promote. Neither is it desirable that the collective of AI experts decide that. From a democratic point of view, we do not want *any* specific group of actors to determine the values that "ethical" technology should promote, but that a plurality of interests in society are allowed to reflect and discuss this in public deliberation. Further, focussing on how to make the technology 'ethical' may not be the right point of departure. An alternative way of approaching unintended side effects with AI would be to take a peaceful, sustainable, and democratic society as the point of departure, and ask where and how AI technology should be used in an ethical society (c.f., Hagendorff 2021). Moreover, considering technological innovation, and the circumstance that what people in the future value may differ from what people value today, any solution needs to be dynamic and adaptable. Whether a technical solution to avoiding harmful side effects is the way to make AI development safe and democratic is nothing that we can envisage, but as the suggestion to somehow build in ethics in the technology is circulating, in theoretical literature as well as among AI experts, it is important to draw attention to challenges with such an approach.

[10] For instance, by explicitly arguing for technical methods to incorporate ethics in AI systems (AI HLEG 2019: 21–22).

10 Some concrete illustrations

Although this is a theoretical paper, discussing expert responsibility on a principal level, it could be fruitful to reflect upon how our theory of forward-looking responsibility would apply in a real-world case. As we did not conduct our own case study for this purpose, we will discuss two events that occurred at the time of writing, one regarding Twitter, and one regarding Facebook.

Twitter has publicly discussed the importance of studying the effects of recommender algorithms on the public conversation, and commissioned a study to understand whether their algorithm may be biased towards a certain political ideology (Jameel 2022; Chowdhury 2021). By analysing millions of tweets, the study (Huszár et al. 2022) compared the amplification—the extent to which a tweet is more likely to be seen—of political content between algorithmically ranked content with content that showed tweets in reverse chronological order, and found that political content was amplified in the algorithmically ranked timelines content. The study also showed that the political right enjoys higher amplification than the political left. While the favouring of the political right may partly be explained by different strategies on Twitter (Huszár 2022: 4),[11] the ranking of certain content higher while reducing the visibility of others is related to the problem of the basis for democratic deliberation that we discuss above. However, from the perspective of forward-looking responsibility, the fact that Twitter announces these results are noteworthy. By admitting that their algorithm may have harmful effects on public conversation, trying to find out more about it, and publishing the findings, Twitter demonstrates forward-looking responsibility for AI development. In contrast, Facebook whistleblower Frances Haugen's testimonies to the US Senate and to the European Parliament in the autumn of 2021 (US Senate 2021a, b; EP 2021a, b) gives a somewhat different picture of how harmful effects of algorithmic recommendations have been met.[12]

Frances Haugen is a computer scientist and former product manager at Facebook, and her focus on algorithmic products and recommender systems makes her an AI expert of the kind we are interested in, but she is also making statements about other AI experts, namely the Facebook leadership. In her testimonies (US Senate 2021a; EP 2021a), Haugen brought up how Facebook has become "a system that amplifies division, extremism and polarisation" and has led to harm for children,

[11] However, the authors point out the precise causal mechanisms behind the favouring of the political right needs to be further investigated (Huszár et al. 2022: 4).

[12] We are aware that a testimony cannot be taken as the truth. However, we do not consider that to be a problem in this case, as we do not draw any conclusions based on it, but only use it as an illustration.

public safety and democracy. What is particularly striking with Haugen's testimonies is that she claims that these are effects of deliberate choices made by the Facebook leadership, that Facebook is fully aware of these effects, and that they have misled "the public, the US Government, its shareholders and governments around the world" about what its research reveals about harmful effects of its recommender algorithms.

We can straightforwardly establish that blowing the whistle was an act of assuming forward-looking responsibility made by an expert who—at the time—was in a position to do so. As a former Facebook employee, she had been witnessing a way of dealing with harmful effects of Facebook's algorithms and recommendations that she found problematic, and she had experienced that attempts to mitigate them from the inside turned out to be futile. In *Time*, she explained that when Facebook dissolved its civic-integrity group, of which she had been a member, she realised that "Facebook no longer wanted to concentrate power in a team whose priority was to put people ahead of profits", and that she needed to leave (Perrigo 2021). In a position from the outside, and with the special expertise that she possessed, Haugen obtained the capacity to act in a way that potentially could affect the future, that is, she was able to take on responsibility in a forward-looking sense. At the same time, she clearly points out other experts, that is, the Facebook leadership, as the ones who really are responsible to do something about the situation.

From a moral perspective, you might be excused from your responsibility to avoid unintended side effects if you are prevented from finding out about these effects. From Haugen's testimonies on how Facebook acts in a misleading way about what they know, one can get the impression that the Facebook leadership tries to avoid responsibility by picturing themselves as being ignorant of the harmful effects of Facebook's recommender algorithm. However, as we argue above, claiming ignorance is not a satisfying way of being excused. On the contrary, you could be responsible to get the knowledge needed, and in the case of the Facebook leadership, as claimed by Haugen, it is not very probable that they are unaware of the risks with the technology they develop. In contrast, according to Haugen's testimonies, they are not only fully aware, but also try to hide it. Further, following our reasoning, by virtue of their expertise regarding technological aspects of AI, AI experts have a larger responsibility than the rest of us to avoid harmful side effects of the AI technology they develop in the first place. If they deliberately hide what they know about how the harmful side effects come about, their responsibility is arguably even larger.

11 Concluding remarks

While the knowledge about technology that AI experts possess makes them better than the rest of us at everything from the potentials of AI technology to how the details in the code should be written to achieve a certain effect, the direction of AI development is not exclusively decided by technological know-how and inventiveness. AI experts do not act in a vacuum. Market demand and the will of funders and political leaders definitely play a part, as does the imagination of science fiction authors and expectations from the public. For sure, AI development is an interplay between a number of different actors from many sectors of society. Nevertheless, it is hard to deny that AI experts have a special position in this process by being the ones with the skills and hands-on capacity to actually make it happen. However, from the capacity to make a difference does not follow a moral right to decide on which difference to make. We certainly do not want AI experts—individually or collectively—to determine which values to promote without listening to other actors in society and taking their views into account. We cannot talk about the forward-looking responsibility of AI experts without reflecting about what this responsibility would imply, and how it relates to the responsibility of other actors.

To conclude, due to their capacity and position, AI experts do have a forward-looking responsibility to influence the direction of AI development, and to some extent, AI experts have a forward-looking responsibility to avoid undesirable, unintended effects of the technology they develop. One could even go a step further and suggest that AI experts have a forward-looking responsibility to investigate the effects of AI technology on society and its individuals, and in order to do that also have a responsibility to acquire a general understanding of the wider context in which their technological solutions will operate. An additional implication of our theory of forward-looking responsibility, that relates to the practical work of AI development, is that AI experts could be responsible for considering unintended harmful side effects of the technology they develop even when their employer or manager has not given them that commission.

However, with this responsibility also comes a risk of increasing the already great influence of AI experts. It is therefore important to pay attention to what the forward-looking responsibility of AI experts should mean in practice. We believe that an important part of the forward-looking responsibility of AI experts is to be as informative as they possibly can when they are involved in policy-making processes and in other contexts in which they are in a position to make an influence. By explaining how the technology works and flagging for potential undesirable side effects, AI experts can raise awareness among decision-makers as well as among the general public. Such an approach contributes to putting critical issues on the political agenda and encourages public debates, which should be the basis for democratic decisions on AI development. If, however, AI experts try to conceal what they know about the origin of unintended effects, as the whistleblower testimony indicates is the case with the Facebook leadership, they do the opposite to what

their responsibility dictates in a forward-looking sense; the Twitter example is more hopeful for the future development of AI in this regard. It is crucial that a diversity of actors is included in democratic processes on the future development of AI, but for this to be meaningful, AI experts need to take responsibility for how the AI technology they develop affects public deliberation.

Author contributions Hedlund contributed knowledge about political and social science theory and main part of the writing. Persson contributed knowledge about ethical and philosophical theory and part of the writing. The ideas and analysis presented in the paper emerged from discussions among the authors where both authors contributed equally.

Funding Open access funding provided by Lund University. This article is written within the research project "How will different forward-looking distributions of responsibility affect the long-term development of Artificial Intelligence?", MMW 2018.0020, funded by Marianne & Marcus Wallenberg foundation.

Data availability Not applicable.

Materials availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

AI HLEG 2019a. *Ethics Guidelines for Trustworthy AI*. European High Level Expert Group on Artificial Intelligence.

AI HLEG 2019b. *Policy and Investment Recommendations for Trustworthy AI*. European High Level Expert Group on Artificial Intelligence.

Baier A (1986) Trust and antitrust. Ethics 96(2):231–260

Ballaert V, Bart, (2015) The politics behind the consultation of expert groups: an instrument to reduce uncertainty or to offset salience? Politics and Governance 3(1):139–150

Bohman J (1999) Democracy as inquiry, inquiry as democratic: Pragmatism, social science, and the cognitive division of labour. Am J Political Sci 43(2):590–607

Boswell C (2009) The political uses of expert knowledge: immigration policy and social research. Cambridge University Press, Cambridge, New York

Chowdhury, Rumman (2021). "Examining algorithmic amplification of political content on Twitter", blogpost, October 21. https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent

Dahl RA (1989) Democracy and Its Critics. Yale University Press

Davis M (2012) 'Ain't no one here but us social forces': Constructing the professional responsibility of engineers. Sci Eng Ethics 18(1):13–34

Dignum, Virginia (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.

Douglas H (2009) Science, Policy, and the Value-free Ideal. University of Pittsburgh Press, Pittsburgh, PA

EC Reg Expert. European Commission *Register of Commission Expert Groups and Other Similar Entities*. High-Level Expert Group on Artificial Intelligence (E03591). https://ec.europa.eu/transparency/expert-groups-register/screen/expert- groups/consult?do=groupDetail.groupDetail&groupID=3%20591

EP (2021a). European Parliament. Whistleblower Frances Haugen testified in European Parliament on November 8, 2021a. https://www.europarl.europa.eu/news/en/press-room/2021a1028IPR16121/facebook-whistleblower-frances-haugen-testifies-in-parliament-on-8-november

EP (2021a). European Parliament. Public Hearing on Whistle-blower's testimony on the negative impact of big tech companies' products on user: opening statement by Frances Hagen, November 8, 2021b. https://multimedia.europarl.europa.eu/fr/video/public-hearing-on-whistle-blowers-testimony-on-the-negative-impact-of-big-tech-companies-products-on-user-frances-haugen-opening-statements_I213108

EC Futurium. European Commission. *AI HLEG – steering group of the European Alliance*. https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai- alliance.html

Feinberg, Joel (1970). "Collective responsibility" in *Collective responsibility: Five decades of debate in theoretical and applied ethics*, Larry May & Stacey Hoffman (eds.), 53–76. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Fossa F (2018) Artificial moral agents: moral mentors or sensible tools? Ethics Inf Technol 20:115–126

Goldman AI (2006) Experts: Which ones should we trust? In: Selinger E, Crease RP (eds) The philosophy of expertise. Columbia University Press, New York, pp 14–38

Gundersen T (2018) Scientists as experts: A distinct role? Stud Hist Philos Sci 69:52–59

Hagendorff, Thilo (2021). "Blind spots in AI ethics", *AI and Ethics*. https://doi.org/10.1007/s43681-021-00122-8.

Haider, Jutta and Olof Sundin (2020). *Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life*. Routledge.

Hansson SO (2012) Safety is an inherently inconsistent concept. Saf Sci 50(7):1522–1527

Hedlund M (2012) "Epigenetic responsibility", *Medicine Studies* 3(2), 171–183, special issue Responsibility in Biomedical Practices. Published Online First. https://doi.org/10.1007/s12376-011-0072-6O

Held V (1970) Can a random collection of individuals be morally responsible? In: May L, Hoffman S (eds) Collective

Responsibility: Five Decades of Debate in Theoretical and Applied Ethics. Rowman & Littlefield Publishers Inc, Lanham, MD, pp 89–100
Hoard S (2015) Gender expertise in public policy: towards a theory of policy success. Palgrave Macmillan, Basingstoke
Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on Twitter. Proc Natl Acad Sci USA 119(1):1–6
Jameel, Shoaib (2022). "Twitter's algorithm favours the political right, a recent study finds", *The Conversation* January 31. https://theconversation.com/twitters-algorithm-favours-the-political-right-a-recent-study-finds-175154
Jasanoff, Sheila (1990). *The Fifth Branch: Science Advisers as Policy Makers.* Harvard University Press.
Jasanoff, Sheila (2016). *The Ethics of Invention*. New York & London: W. W. Norton & Company.
Kant, Immanuel (1999) *Critique of Pure Reason*. Cambridge University Press. [orig. *Kritik der reinen Vernunft* Johann Friedrich Hartknoch verlag 1781]
Knaggård, Åsa, Erik Persson, Kerstin Eriksson (2020). "Sustainable distribution of responsibility for climate change adaptation". *Challenges* 11(1).
Kudina O, Verbeek P-P (2018) Ethics from within: Google glass, the Collingridge dilemma, and the mediated value of privacy. Sci Technol Human Values 44(2):1–24
Lang Jr., Anthony F. (2015). "Shared political responsibility" in *Distribution of Responsibilities in International Law*, André Nollkaemper and Dov Jacobs (eds.). Cambridge: Cambridge University Press, pp. 62–86
Lidskog R, Sjödin D (2018) Unintended consequences and risk(y) thinking: the shaping of consequences and responsibilities in relation to environmental disasters. Sustainability 10(8):2906–2922
Lindvall J (2009) The real but limited influence of expert ideas. World Politics 61(4):703–730
Manin B, Stein E, Mansbridge J (1987) On legitimacy and political deliberation. Political Theory 15(3):338–368
Matzner N, Barben D (2020) Climate engineering as a communication challenge: contested notions of responsibility across expert arenas of science and policy. Sci Commun 42(1):61–89
Metz J (2015) The European Commission, expert groups and the policy process: demystifying technocratic governance. Palgrave Macmillan, Basingstoke
Metzinger, Thomas (2019). "EU guidelines: Ethics washing made in Europe", *Der Tagespiegel*, April 8.
Miller S (2001) Collective responsibility. Public Aff Q 15(1):65–82
Muyskens JL (1982) Collective responsibility of the nursing profession. In: May L, Hoffman S (eds) Collective responsibility: five decades of debate in theoretical and applied ethics. Rowman & Littlefield Publishers Inc, Lanham, MD, pp 167–178
Orr W, Davis JL (2020) Attributions of ethical responsibility by Artificial Intelligence practitioners. Inf Commun Soc 23(5):719–735
Perrigo, Billy (2021). "How Facebook forced a reckoning by shutting down the team that put people ahead of profits", *Time,* October 7, 2021. https://time.com/6104899/facebook-reckoning-frances-haugen/
Persson E, Hedlund M (2021) The future of AI in our hands? To what extent are we as individuals morally responsible for guiding the development of AI in a desirable direction? AI Ethics. https://doi.org/10.1007/s43681-021-00125-5
Persson E, Eriksson K, Knaggård Å (2021) A fair distribution of responsibility for climate adaptation—translating principles of distribution from an international to a local context. Philosophies 6(3):68
Scheffler S (2018) Why Worry About Future Generations? Oxford University Press
Schraube E (2009) Technology as materialized and its ambivalences. Theory Psychol 19(2):296–312
Schudson M (2006) The trouble with experts—and why democracies need them. Theory Soc 35(5/6):491–506
Sirbu A, Pedreschi D, Giannotti F, Kertész J (2019) Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. PLoS ONE 14(3):1–20
US Senate (2021a). United States Senate. Committee on Commerce, Science, & Transportation, Subcommitte on Consumer Protection, Product Safety, and Data Security. "Protecting Kids Online: Testimony from a Facebook Whistleblower", October 5, 2021a. https://www.commerce.senate.gov/2021a/10/protecting kids online: testimony from a facebook whistleblower
US Senate (2021a). United States Senate. Committee on Commerce, Science, & Transportation, Subcommittee on Consumer Protection, Product Safety, and Data Security. *Statement of Frances Haugen,* October 4, 2021b.
Van de Poel, Ibo (2015a). "Moral responsibility" in *Moral Responsibility and the Problem of Many Hands,* Ibo van de Poel, Lambèr Royakkers, and Sjoerd D. Zwarf (eds.). London & New York: Routledge, pp. 12–49.
Van de Poel, Ibo (2015b). "The problem of many hands" in *Moral Responsibility and the Problem of Many Hands,* Ibo van de Poel, Lambèr Royakkers, and Sjoerd D. Zwarf (eds.). London & New York: Routledge, pp. 50–92.
Van de Poel, Ibo, Lambèr Royakkers, and Sjoerd D. Zwarf (2015). *Moral Responsibility and the Problem of Many Hands*. London & New York: Routledge.
Verbeek P-P (2006) Materializing Morality: Design Ethics and Technological Mediation. Sci Technol Human Values 31(3):361–380
Walsh T (2018) Machines that Think: The Future of Artificial Intelligence. Prometheus Books, New York
Watson JC (2019) What experts could not be. Soc Epistemol 33(1):74–87
Young IM (2006) Responsibility and global justice: A social connection model. Social Philosophy and Policy Foundation 23(1):102–130
Young IM (2011) Responsibility for Justice. Oxford University Press, New York