Philosophy & Technology (2021) 34:1057–1084
https://doi.org/10.1007/s13347-021-00450-x

RESEARCH ARTICLE

Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them

Filippo Santoni de Sio[1] · Giulio Mecacci[2]

Received: 26 May 2020 / Accepted: 25 March 2021 / Published online: 14 May 2021

Abstract

The notion of "responsibility gap" with artificial intelligence (AI) was originally introduced in the philosophical debate to indicate the concern that "learning automata" may make more difficult or impossible to attribute moral culpability to persons for untoward events. Building on literature in moral and legal philosophy, and ethics of technology, the paper proposes a broader and more comprehensive analysis of the responsibility gap. The responsibility gap, it is argued, is not one problem but a set of at least four interconnected problems – gaps in culpability, moral and public accountability, active responsibility—caused by different sources, some technical, other organisational, legal, ethical, and societal. Responsibility gaps may also happen with non-learning systems. The paper clarifies which aspect of AI may cause which gap in which form of responsibility, and why each of these gaps matter. It proposes a critical review of partial and non-satisfactory attempts to address the responsibility gap: those which present it as a new and intractable problem ("fatalism"), those which dismiss it as a false problem ("deflationism"), and those which reduce it to only one of its dimensions or sources and/or present it as a problem that can be solved by simply introducing new technical and/or legal tools ("solutionism"). The paper also outlines a more comprehensive approach to address the responsibility gaps with AI in their entirety, based on the idea of designing socio-technical systems for "meaningful human control", that is systems aligned with the relevant human reasons and capacities.

Keywords AI ethics · Responsibility gap · AI and accountability · Meaningful human control · Engineer responsibility

✉ Filippo Santoni de Sio
f.santonidesio@tudelft.nl

[1] Section Ethics/Philosophy of Technology, Delft University of Technology, Delft, The Netherlands

[2] Department of Cognitive Artificial Intelligence, Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

1 Introduction

In 2004, Andreas Matthias introduced what he called the problem of "responsibility gap" with "learning automata" (Matthias, 2004). In a nutshell, intelligent systems equipped with the ability to learn from the interaction with other agents and the environment will make human control and prediction over their behaviour very difficult if not impossible, but human responsibility requires knowledge and control. Therefore, we as humanity are facing a dilemma: either we go on with the design and use of learning systems, thereby giving up on the possibility of having human persons responsible for their behaviour, or we preserve human responsibility, and thereby give up on the introduction of learning systems in society. Matthias formulation of the responsibility gap has been quite influential especially in relation to the development of autonomous weapon systems (Sparrow, 2007)(Human Right Watch, 2015).

More recently, the concern with "responsibility gaps" has been raised more generally in relation to artificial intelligence (AI), that is to any technique designed to solve problems traditionally assigned to human intelligence (Amoroso & Tamburrini, 2019). Risks of gaps have been identified in relation not only to the learning capacities of AI but first and foremost to the opacity, complexity, and unpredictability that these systems generally display (Mittelstadt et al., 2016). In fact, the question so as to what extent persons can or should maintain responsibility for the behaviour of AI has become one of, if not the most discussed question in the growing field of so-called ethics of AI (Braun et al., 2020; Coeckelbergh, 2019; Nyholm, 2018).

However, whereas "responsibility" is known in philosophy and law for being an ambiguous and polysemantic term (Hart, 1968; Feinberg, 1970), this complexity is rarely reflected in the debates on responsibility for the behaviour of systems that include AI. Therefore, discussions about "responsibility" or "accountability gaps" are sometimes partial as they usually appeal to a non-sufficiently specified notion of responsibility. Moreover, the focus on learning automata or "autonomous systems" may be too limited. Responsibility gaps are due to a multiplicity of factors and are sometimes only aggravated by the presence of machines that learn and act on their own.[1] In fact, sufficiently interconnected socio-technical systems, with limited artificial intelligence and capacity to learn, but relying on a complex texture of human agents and technical systems, such as bureaucracies or corporates, might also generate responsibility gaps. Considering different causes of responsibility gaps related to AI and automation beyond "autonomy" and "learning" will contribute to carve the problem at the right joints and provide better insights towards possible solutions.

The first goal of this paper is thus reframing the responsibility gap discussion in terms that are better aligned with the categorisation of responsibility concepts in moral and legal philosophy. By doing so, we will be able to better see *what kind of*

[1] One important additional reason to shift the focus away from "autonomous systems" is that the concept of machine autonomy itself is very controversial, tends to raise philosophical debates that do not help clarifying the issue of responsibility gaps, and may shift attention away from other urgent ethical issues with (non-autonomous) AI.

responsibility is threatened *by which aspect of automation* and *why this matters*.[2] We take some concepts and distinctions from philosophical, legal, and sociological theory of responsibility (Bovens, 1998; Collingridge, 1980; Hart, 1968; Pesch, 2015; van de Poel & Sand, 2018), and we use them to identify four different kinds of responsibility gaps: the *culpability* gap, the *moral accountability* gap, the *public accountability* gap, and the *active responsibility* gap. We will also identify, for each of those, different possible causes, integrating Matthias' classic analysis that identified learning capacities of automata as the main source of responsibility gaps, generically speaking. To a deeper look, Matthias addressed in his work what we will call the culpability gap – the risk that no human agent might be legitimately blamed or held culpable for the unwanted outcomes of actions mediated by AI systems. Gaps in this kind of responsibility have already received some attention, both from a moral (Matthias, 2004; Sparrow, 2007) and a legal perspective (Calo, 2015; Pagallo, 2013). Attention also went to the "accountability gap" in relation to autonomous weapon systems (AWS) (Heyns, 2013); (Meloni, 2016), and more generally within the discussion on explainability of algorithms and AI (Mittelstadt 2016; Doran et al. 2017; Pasquale, 2016). However, we propose to distinguish two forms of the accountability gap: the "public accountability gap", i.e. citizens not being able to get an explanation for decisions taken by public agencies, and a broader "moral accountability gap" – i.e. the reduction of human agents' capacity to make sense of – and explain to each other the "logic" of their behaviour, due to the mediation of opaque, unexplainable algorithms and complex autonomous systems and/or the lack of appropriate psychological, social incentives or institutional spaces that promote these explanations. One particularly important form of the moral accountability gap is that concerning the difficulty for engineers and other agents involved in the process of technological development to systematically discuss with one another their understanding of the goal and meaning of this process. Finally, the "active responsibility gap" has not to our knowledge been addressed as such so far. This gap consists in the risk that persons designing, using, and interacting with AI may not be sufficiently aware, capable, and motivated to see and act according to their moral obligations towards the behaviour of the systems they design, control, or use. In particular,this gap concerns the obligation to ensure that these systems do not impact negatively on the rights and interests of other persons and, ideally, positively contribute to their well-being instead. Distinguishing four responsibility gaps and their various sources is the focus of the first part of this paper.

In the second part, we will show some of the common approaches that have so far been taken towards addressing responsibility gaps. We show how those approaches offer a partial and limited understanding of the responsibility gap and thus offer solutions that would apply only to specific aspects of them. Those who will be here called "fatalists" (Matthias, 2004; Sparrow, 2007) tend to focus on a too limited understanding of the responsibility gap, that is a gap in *culpability* for the behaviour of *learning* technological automata. Those who will be here called "deflationists" (Simpson & Müller, 2016) underestimate the novelty of the AI revolution and its implication for culpability attributions in morality (blameworthiness) and the law (liability); they also

[2] See (Di Nucci & Santoni de Sio, 2016) for an early attempt to clarify this.

seem to underestimate the risks of gaps in the moral and political accountability of system designers as well as gaps in theirs and other agents' active responsibility for the behaviour of artificial intelligence. Promoters of "explainable AI" and other scientific and technological improvements tend to ignore the psychological, social, and political dimension of the interaction with AI, thereby running the risk of embracing some form of "technical solutionism" (Stilgoe, 2017), by which all the moral and social problem of human responsibility for the behaviour of artificial intelligence can be fixed simply by an improvement of the working of AI techniques. Lawyers and policy-makers proposing the revision of current legal liability regimes (including extension of strict and product liability regimes, and "electronic personhood") may either underestimate the importance of maintaining some form of human *moral* responsibility on the behaviour of the artificial intelligence or recognise this need but without saying how moral and social practices – and not only legal rules – should change in order to govern a responsible transition to the use of AI. We call this the risk of "legal solutionism". One result of this critical review is the recognition that the different notions of responsibility, though distinct, are also interconnected and that often addressing one kind of gap requires attention to one or more of the others. This suggests the necessity of a more integrated and comprehensive approach.

We will conclude this paper by sketching such a more encompassing approach which, as we argue, can contribute to address a larger number of gaps in their interconnections. We will suggest that one recent approach to "meaningful human control" (MHC) (Mecacci & Santoni de Sio, 2020; Santoni de Sio & van den Hoven, 2018) might be suitable to frame the several responsibility gaps within a bigger scheme and offer principles to transversally address them. Future research will develop this proposal in more detail.

2 Varieties of Responsibility Gaps

The term "responsibility" has different meanings. H.L.A. Hart's classical account (Hart, 1968) lists four senses (role-responsibility, causal responsibility, capacity-responsibility, liability-responsibility). Based among others on the work of John Gardner (2007), Mark Bovens (1998), and Ibo van de Poel (2015), we work with a revised and modified list of four forms of responsibility that are particularly relevant – yet not limited to – the context of automation and artificial intelligence: culpability, moral accountability, public accountability, and active responsibility. The next four sub-sections present these four forms of responsibility and identify the specific related challenges presented, or furthered, by the introduction of artificial intelligence (Table 1). Some examples will serve as illustration.

2.1 Culpability Gaps

When things go wrong and important interests or rights such as physical integrity or life are infringed, we, as victims and as society, not only want to understand what happened and why. We also want to know whether the harm was the result

Table 1 Types of responsibility gaps

Type of responsibility	Definition	Gaps with AI
Culpability	Blameworthiness for wrongdoing based on intention, knowledge or control	AI making prediction and control more difficult, thereby creating new legitimate reasons/excuses for wrongdoing, e.g. an avoidable road crash involving an automated driving system that nobody could individually predict or prevent
Moral accountability	Duty of human persons to explain one's reasons and actions to others (under some circumstances)	AI making processes unexplainable to the very persons using it, e.g. a doctor not being able to explain the reasons for her diagnosis to a patient
Public accountability	Duty of public agents to explain their actions to a public forum	AI shifting discretionary powers towards IT experts and data analysts (often outsourced to private companies) whose work is harder to publicly scrutinise, e.g. government using (private) AI-systems in support of their decision-making
Active responsibility	Duty to promote and achieve certain societally shared goals and values	Actors involved in the design or use of AI not being sufficiently aware of their own responsibility to prevent harm deriving from AI or not being able or motivated to fulfil this obligation, e.g. engineers or managers only looking at the technical benefits of AI

of someone's wrong behaviour, and if it turns out that the wrong behaviour is one for which there is no justification or excuse, we want the author to be condemned, sanctioned, or even punished for their behaviour. This is in a nutshell culpability or blameworthiness. Whether and to what extent attributions of culpability make sense has been the main subject of the centuries-long philosophical debate on free will, typically in the light of causal determinism (is it fair to blame each other if all our actions are necessarily caused by previous physical and mental events?) or more generally in the light of a view of human behaviour shaped by (neuro)scientific knowledge (if human action can be fully explained by behavioural/social/neuroscience, what is left of moral culpability?) (for a general discussion along these lines, see, e.g. (Pereboom, 2006)). However, many philosophers and most lawyers and laypersons do believe that (fair) social and legal practices of attribution of culpability should also be maintained and promoted (as opposed to just relying on any form of social and psychological education or therapy) (Morse, 2006), at least to the extent they are the legitimate expression of appropriate moral sentiments by the wronged individuals and society at large (Strawson, 1962), they reinforce the social commitments to shared norms (Sie, 2005), and, possibly most importantly, they contribute to control and reduce undesirable behaviour. Similarly, also state-administered punishment for serious criminal behaviour may be morally defensible and even desirable insofar as it gives effectiveness to the expression of public condemnation (Feinberg, 1965), and serves the utilitarian goals of discouraging similar behaviour by the defendant themselves in the future and by other citizens more generally. Finally, (public) attributions of culpability have the function to compensate the victims – symbolically, or even materially, typically in the case of compensations to plaintiffs in civil litigations.

AI-based systems may put culpability practices under stress in different ways, preventing the realisation of one or more of their goals. Consider, as an example, automated driving systems (ADS). First, ADS may make the network of agents involved in the driving more complex, just because more agents are involved and/or new forms of interactions are created. For instance, a vehicle may be operated by a driver D1, with the assistance of the automated driving system AS, produced by the car manufacturer X, powered with digital systems developed by the company Y, possibly including some form of machine learning developed by the company Z, and enriched by data coming from different sources, including the driving experience of drivers D2, D3…Dn; vehicles in this system are in principle subject to a standardisation process done by the agency S, the traffic is regulated by the governmental agency G, drivers are trained and licensed by the agency L etc. Second, some specific features of present-day learning AI systems may make this interaction particularly unpredictable – typically when the vehicles' performance is potentially re-designed by the second on the basis of new data acquisition and processing – and opaque, if the reasoning scheme underlying systems' actions is not easily accessible to their controllers, regulators, or even their designers.

Agents operating in such a socio-technical system (designers, programmers, drivers, regulators, bystanders etc.) may more easily find themselves acting wrongly, for instance, causing an avoidable road crash, while at the same time having a legitimate excuse. Nobody, and certainly not them, could reasonably predict certain circumstances or reasonably avoid a certain outcome, therefore not being open to legitimate blame (Matthias, 2004; Sparrow, 2007). We call this the "culpability gap".

The culpability gap has not been created by the introduction of "learning automata" (machine learning) and their inherent unpredictability, as it has been framed by some authors (e.g. Matthias, 2004). As a matter of fact, other intelligent, autonomous entities with "no soul to blame and no body to kick" (Asaro, 2012), such as, e.g. bureaucracies and corporates, may in themselves generate gaps in culpability.[3] This has been classically identified as "the problem of many hands" (Bovens, 1998; Thompson, 1980). Artificial intelligence plays however an important role by contributing to create new versions of the phenomenon and thus making it more visible. Also, the use of artificial intelligence and data-driven machine learning in decision-making importantly introduces a new element of technical opacity and lack of explainability that makes it more difficult for individual persons to satisfy the traditional conditions for moral and legal culpability: intention, foreseeability, and control.[4]

Culpability gaps are concerning insofar as the more persons designing, regulating, and operating the system can legitimately (and possibly systematically) avoid blame for their wrong behaviour, the less these agents will be incentivised to prevent these wrong behaviours. In fact, they will arguably have less incentives to strive for a high(er) level of safety, awareness, attention, motivation, and skilfulness. Also,victims of unjust harm will be less likely to receive compensation. Finally, it might become more difficult for persons more generally to make sense of their moral sentiments in relation to wrongs and accidents and to direct one's reactive attitudes towards some legitimate target. This may impoverish the human capacity to express moral judgement and may feed helplessness and moral scepticism towards the possibility of understanding and rectifying wrongdoing. As noted by Danaher (2016), the desire to find a scapegoat to satisfy these feelings may also be fuelled.

2.2 Moral Accountability Gaps

Culpability is a particularly serious form of (moral) responsibility, but it is not the only one. Individual persons are often called to respond for their choices or actions in less threatening ways, for instance, when family, friends, or acquaintances ask them why-questions, not necessarily with the intention of judging or blaming them but possibly to just engage in a conversation and to better understand each other's reasons and expectations. Why were you late at the appointment, why did you start taking guitar classes, why did you turn down that job offer…? We will call this expectation to answer (at least some) why-questions from other persons' "moral accountability", to distinguish it from the "public accountability" discussed below. Moral accountability has been presented in the philosophical literature on

[3] As a broader philosophical point, it has been argued that even cultural systems can be seen as types of "automata" (Sini, 2021) for their evolution cannot be controlled by individuals while rather shaping and constraining individual behaviour.

[4] More on this in the section "Fatalism and deflationism" below.

moral responsibility as a key-aspect for the justification and understanding of moral responsibility practices (Wolf, 1990; McKenna, 2012). The legal philosopher John Gardner calls it "basic responsibility" insofar as he sees it as the core of what it means to be a reflective person in society (Gardner, 2007). In this sense, being accountable, unlike being culpable, is something to be desired rather than avoided insofar as it is a constitutive part of being able to reflect on one's actions and to participate in meaningful social relations. It also helps persons seeing events in the world as connected to their rational capacities and thereby supporting their sense of agency and responsibility (Honoré, 1999). It is a classic view of human responsibility, which can be reported back to the old Socratic motto "know thyself". In fact, Gardner calls this the "Aristotelian story" about responsibility, insofar as it focuses on the persons' capacity to *make sense of* theirs and others' actions and choices as something connected to their abstract reasons (as opposed, for instance, to just physical or biological events).

Moral accountability also has an instrumental value. The process of exchanging questions and reasons helps finding explanations for things that have happened, reinforces trust and social connections between agents, and by exposing persons to potential requests for explanation and justification, it also tends to reduce undesired behaviour by pushing persons to be more clearly aware of the impact of their actions on others and therefore motivated to prevent unwanted outcomes (and potential blameworthiness). In relation to engineering practices, Genus and Stirling (2018) have recently stressed out the importance of Collingridge's proposal to "engage more strongly with accountability in debates bearing on key elements of responsible innovation" (Collingridge, 1980, p. 62). In their view, also inspired by Lindblom (1990), accountability is a key tool to enhance the reflexivity of the agents involved in the design and development of (new) technologies, and to promote responsiveness between them and those who will be affected by their creations. In the literature on "responsible research and innovation", the importance of accountability practices has been recently emphasised and categorised under the label of "responsiveness between stakeholders" (Stilgoe et al., 2013).

Moral accountability may be blurred in different ways by the introduction of artificial intelligence. First, in general, similar to what observed above about culpability, by contributing to create a more complex chain of decision-making and action, AI may make more difficult for individual agents to make sense of the reasons why a certain decision was taken, what their role exactly was in the operation, and, in general, whose reasons and what reasoning were governing the system they are part of. However, data-driven machine (deep) learning, due to its intrinsic opacity, might make a system's behaviour extra hard to understand and explain. In addition, the whole process of technology development and production is arguably pervaded by an increasing pressure towards deploying proprietary technologies that, even when working through mechanisms accessible to their developers and programmers, are designed to be inaccessible to public scrutiny and the users themselves (Pasquale, 2015). Technology developers, driven by both the desire to minimise industrial espionage and maximise customers loyalty (by, e.g. binding them to their assistance network), will usually avoid sharing data and engineering insights.

An example of AI affecting moral accountability specifically due to the opacity and complexity of AI may be that of a medical doctor using an AI-driven system for diagnosing. These systems are usually based on deep learning techniques that require a thorough training over a dataset the nature of which is well-known and clear. In other words, the system will train on a set of well-known, well-established cases, before being applied to new and unknown cases. The way knowledge is represented in the machine and the exact way the machine distinguishes between a positive and negative diagnosis are not only inaccessible to the doctor who uses the system but also, to an important extent, to those who designed it (Castelvecchi, 2016). Therefore, the capacity of the different agents, including the users, to make sense of the "logic" of the systems' behaviour may be weakened and sometimes lost, together with their capacity and willingness to engage in a meaningful conversation about their role and the responsibility that comes with it. This may create different kinds of problems, depending on the (professional) context, and the roles, the responsibilities, and the human and social relations pertaining to it. The general concern is that AI may make individual persons less able to understand, explain, and reflect upon their own and other agents' behaviour. Let us call this the *moral accountability gap*.

2.3 Public Accountability Gaps

One specific form of accountability is attached on politicians, civil servants, and other agents invested with a public function: public accountability. Public accountability,[5] in Mark Bovens' definition, is a "relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences" (Bovens, 2007). According to Bovens, effective mechanisms of accountability may enhance both the effectiveness of a complex public decision-making system and its compliance with the liberal-democratic values (Bovens, 1998, 2007). Accountability promotes democratic control (transparency), limits abuses of power (corruption), but also brings more effectiveness in the institutions. In fact, providing administrators with information about their own functioning and forcing them to reflect on their successes and failures will eventually allow and encourage them and others to improve their future performances.

There has recently been a big legal and political debate on the extent to which the introduction of AI-based automated (administrative) decision-making is desirable and legitimate (European Commission, 2019; Hildebrandt, 2019; Noto La Diega, 2018). Also, it has been doubted that the GDPR provisions guarantee sufficient transparency and accessibility to these mechanisms for those who are subject to them (Edwards & Veale, 2017; Wachter et al., 2017). Most of these discussions point to the fact that algorithmic decision-making is often difficult to understand for – and explain to – human agents, due to the different, and sometimes inscrutable for persons, ways of AI operations: the so-called "black box" problem (Castelvecchi, 2016). However, Noto

[5] Bovens just speaks of "accountability", but it is clear from the context of his work that he is talking of public accountability; we make the specification explicit to distinguish this from "moral accountability' discussed above.

La Diega (2018) correctly notices that issues of explainability may arise not only due to technical black boxes but also due to what he calls organisational and legal black boxes created or aggravated by the introduction of AI in public administration.

Zouridis et al. (2019) have further explained the sources of these organisational and legal black boxes and their relevance to public accountability. Traditionally, public agencies were organised as "street-level" bureaucracies. Processes were managed by individual case managers who had direct contact with individual citizens and substantial discretionary powers (Bovens & Zouridis, 2002; Lipsky, 1980). With the introduction of digital decision-making systems, these discretionary powers of the street-level professionals have been disciplined. However, this has also greatly shifted the locus of administrative discretion from individual public officers to IT experts, responsible for programming the decision-making process and translating the legislation into software, and to the data analysts, who are responsible for the acquisition and analysis of data (Zouridis et al., 2019). Moreover, these "system-level" bureaucracies are part of larger networks and chains of delegation in which data are exchanged and reused (Van Eck, 2018). This shift raises three challenges for public accountability. First, development in information technologies is often outsourced to private parties or tech-giants, such as Google, who are not politically accountable and may not be willing to disclose critical information about the functioning of their systems (Pasquale, 2015). For example, in her book *Automating Inequality*, Virginia Eubanks (2018) tells the stories of private contractors not being able and/or willing of disclosing the reasons for the failures/mistakes in their digital systems used by some US states for welfare benefit allocation procedures. Second, more generally, the work of the software engineers and data professionals in public organisations is usually not visible and subject to public and legal scrutiny. Finally, far more data are exchanged between many different (public) organisations than in the past. In this way, the introduction of AI makes the "problem of many hands" (Bovens, 1998; Thompson, 1980) more acute: data coming from different sources are introduced and enriched at different points in the data chain. Individual citizens may have a hard time finding out who they should turn to, if data are incorrect, corrupted, or biased as a collective outcome of a series of minor contributions. Technical, organisational, and legal black boxes are the sources of what we call the *public accountability gap* with artificial intelligence.

2.4 Active Responsibility Gap

The philosophical literature on professional responsibility of engineers usually distinguishes between "active" and "passive" responsibility (Bovens, 1998). In a nutshell, active responsibility is forward-looking and concerns the goals, values, and (legal) norms that professionals such as engineers are supposed to promote and comply with as well as the consequences they need to prevent and avoid.[6] Passive responsibility is backward-looking and concerns the moral and legal consequences engineers must face in case something goes wrong. The three forms of moral

[6] Not to be confused with the "forward-looking account of culpability" in the debate on the justification of practices of attribution of moral blameworthiness.

responsibility discussed above – culpability, moral, and public accountability – are all forms of passive responsibility. The legal duty to provide high standards of safety and the so-called corporate social responsibility of companies would be typical examples of active responsibility.

One well-known problem with the active responsibility of engineers is that while engineers have arguably an individual active responsibility to promote societal goods, their work is most often embedded in networks of different agents and institutions (Swierstra & Jelsma, 2006). They may, for instance, be involved in projects connecting scientists and their academic institutions and technological companies operating on the market. As seen above, this may create problems for the attribution of passive responsibility, as in case something goes wrong, it may be easy (and sometimes legitimate) for individual engineers to shift responsibility to other agents or institutions in their network, and it sometimes may be even the case that nobody can legitimately be held responsible for one specific unwanted event (Van de Poel et al., 2015). What is often overlooked is that the networked nature of the engineering work may also create issues with the attribution of active responsibility. As noted by Pesch (2015), engineers may not have a clear and consistent representation of what their (social) role is – are they scientists, technicians, business persons? what are the goals and values they should strive for: truth, innovation, market shares? They may not even have clear and shared systems of principles, norms and rules to follow in their profession, and/or the capacity or motivation to reflect upon and interpret these rules in concrete cases.[7]

Based on the general framework above, we propose that the introduction of AI may create two different but related sets of issues. First, engineers and other agents involved in the development and use of technology may not be (fully)*aware* of their respective moral and social obligations towards other agents. Think, as an example, of a manager of an IT company who, as a result of her personal education or the engineering and business culture in which she has been raised, is genuinely and sincerely convinced that: (a) she is benefitting the public by providing them with more comfort through the use of their new products and that (b) it is not her responsibility to try to minimise the possible negative impact of the use of these products on people's well-being, privacy, or political freedom.[8] In Van de Poel's and Sand's (2018) classification of active responsibility, this is a gap in "obligation". Second, engineers and other agents involved in the development and use of technology may

[7] According to Pesch (Pesch, 2015), what we call "the active responsibility gap" of engineers has old historical roots. Institutions like the modern state, the market, and science, have since the time of their birth been the object of systematic intellectual reflection by statesmen, scientists, philosophers, etc., and they were therefore seen and built according to a complex "intellectual template" ((Pesch, 2015),p 930). Modern technology development, on the other hand, "has never been integrated into this template. Technology has never received the same intellectual attention as a foundation for modern life as the realms of politics, economics, and science, and its integration into these realms has been largely a contingent one" ((Pesch, 2015), p 930). In fact, whereas political philosophy, philosophy of science and political economy are well-established disciplines with deep historical roots; philosophy of engineering is in comparison a relatively new and not even fully established and recognised academic discipline. This problem concerns primarily engineers, but, as we will see more in detail below, also other professionals involved in the design, development, regulation, control, and even use of technology.

[8] A remarkable example of this attitude is that of the IT developers who invented popular social network platforms such as Facebook and Twitter, interviewed in the Netflix documentary *The Social Dilemma*.

not be *sufficiently able or motivated* to fulfil an obligation they may be well aware of. Think, as an example, of military personnel using a new AI-based weapon system: while being perfectly aware of their general obligation to use the system in compliance with the requirement of international law, they may end up making an illegal use of the system, due to insufficient technical training, and/or not having (yet) been able to develop a sufficient capacity to resist the pressure to use the technology in a certain way, coming from superiors and her environment more generally. In van de Poel and Sand's (2018) classification of active responsibility, this is a gap in "virtue", i.e. the concrete capacity and inclination to perform according to certain norms and principles. Let's call these two issues, taken together, the *active responsibility gap*.

3 Partial Answers to Responsibility Gaps: "Fatalism", "Deflationism", and the Risks of "Solutionism"

In the previous section, we have seen how considering different senses of responsibility allows to highlight the existence of four different kinds of responsibility gaps with AI (see Table 1). The problem of responsibility gap with AI, as it turns out, is not one problem but a set of at least four interconnected problems – gaps in culpability, moral accountability, public accountability, and active responsibility. Moreover, these gaps are caused by different sources, some of which are old, i.e. the complexity of social and technical systems; some new, i.e. the data-driven learning features of present-day AI; some more technical, i.e. the intrinsic opacity of algorithimic decision-making; some more political and economic, i.e. the implicit privatisation of public administration; and some more societal, i.e. the engineers' and other actors' lack of awareness and/or capacity to comply with their (new) moral, legal, societal obligations. Sufficient awareness of this complexity has been missing in the debate so far. Current debates have posed a strong accent, from time to time, on mainly one of these different problems and one or two of their sources thereby often not giving sufficient attention to the broader picture. In this second part of the paper, we make a critical revision of a representative sample of the literature on the responsibility gap, with a twofold goal: first, to show the extent to which this literature misses out on the complexity of the responsibility gap, and second, to suggest that taking such a partial approach to the responsibility gap not only brings the risk of offering an incomplete picture of the problem but also a distorted one. We present three possible distortions in the presentation of the responsibility gap with AI, which we call, respectively: "fatalism", i.e. the idea that the responsibility gap is a new and intractable problem; "deflationism", i.e. the idea that the responsibility gap is not new and not a problem; and "solutionism", i.e. the idea that the responsibility gap is a problem that can be solved by simply introducing new technical and/or legal tools (Fig. 1). To be clear, we are not claiming that every author discussing one aspect of the responsibility gap should necessarily also be victim of one of these distortions: as a matter of fact, many authors are well aware that they are addressing only one aspect of a more complex problem. What we are suggesting is that, in the long run, focusing only on these partial analyses may provide a misleading picture of the

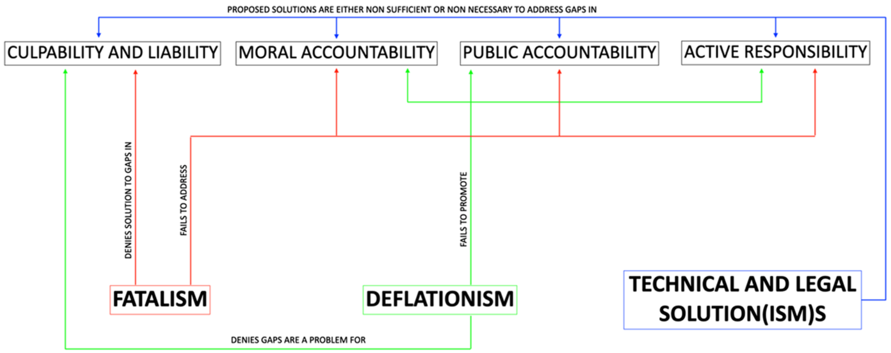
PROPOSED SOLUTIONS ARE EITHER NON SUFFICIENT OR NON NECESSARY TO ADDRESS GAPS IN
CULPABILITY AND LIABILITY
MORAL ACCOUNTABILITY
PUBLIC ACCOUNTABILITY
ACTIVE RESPONSIBILITY
DENIES SOLUTION TO GAPS IN
FAILS TO ADDRESS
FAILS TO PROMOTE
FATALISM
DEFLATIONISM
TECHNICAL AND LEGAL SOLUTION(ISM)S
DENIES GAPS ARE A PROBLEM FOR

Fig. 1 Partial answers to responsibility gaps and their limits

responsibility gap as well as hindering the creation of an appropriate response. In the last part, we will preliminarily consider how a more encompassing approach, such as designing for "meaningful human control", might represent such an alternative, more appropriate answer.

3.1 Fatalism and Deflationism

At the two extremes of the debate on responsibility gaps are those whom we will call, respectively, fatalists and deflationists. The fatalist approach is best captured by Matthias, 2004 paper on "The responsibility gap for learning automata". According to Matthias, the introduction of learning automata in society poses an unprecedented challenge to moral responsibility (culpability) and presents us with a moral dilemma. Culpability requires knowledge; learning systems make knowledge (i.e. prediction of outcomes) impossible; therefore, learning automata make impossible to (legitimately) attribute culpability to human persons for the actions mediated by learning systems. We as a society are then facing a dilemma: either we introduce learning systems and give up on culpability, or we maintain culpability, and give up on the introduction of learning systems in society (Matthias, 2004). Matthias formulation of the responsibility gap has been quite influential especially in relation to the development of autonomous weapon systems (Sparrow, 2007). In this perspective, the responsibility gap created by learning automata (AI) is a new, serious, and intractable problem. We have argued above that the culpability gap is not completely new, and we will suggest below how it could be potentially addressed by designing socio-technical systems with a new notion of human control.

At the other extreme of the debate are those who believe that the culpability gap for learning automata and AI more generally is not a problem after all, and anyway not a new one, so that old technical, moral, and legal recipes will just suffice. We will call these "deflationists". Some deflationists just embrace the first horn of Matthias' dilemma. If we have reasons to believe that the introduction of learning automata will bring significant societal benefits, for instance, in terms of efficiency, effectiveness, or overall safety of the processes, we may and should introduce these in society, even if this will lead to some erosion of human moral responsibility (in any possible sense), for the (*ex hypothesi* reduced in number) accidents. For instance, many believe that, if the introduction of AI in critical tasks such as medicine, transport, and warfare is likely to reduce the overall number of deaths or injuries by reducing the impact of human error, then we should not care too much about the risk of gaps in moral responsibility. This is a too simple utilitarian approach, which tends to obfuscate the moral relevance of individual moral and legal duties and rights (Santoni de Sio, 2017). It also ignores the moral relevance of fairness and distributive justice in assessing the moral risks of technology (Hayenhjelm & Wolff, 2012). However, Simpson and Müller (2016) have proposed a more sophisticated version of deflationism, one which embraces the first horn of Matthias dilemma (accepting learning systems *and* some gaps in moral culpability), while trying to pay due respect to individual rights and distributive justice. Simpson and Müller admit that AI may create some culpability gaps, but, so they claim, this is not new. Also, non-intelligent, non-learning systems like bridges and buildings sometimes collapse (fail) without any human person being culpable for that, and we as a society find this acceptable, insofar as all reasonable precautions had been taken, and nobody could have reasonably prevented the accident. The same can be said about AI. In contrast to utilitarian approaches, Simpson and Müller take a "non-aggregative" and "contractarian" approach to the ethics of risk and claim that we as a society have two obligations: to reduce the aggregate risks involved in the use of technology, and to minimise the risk of harm to *each* of the persons involved. If these two goals are achieved by the introduction of a given technology, and a "tolerance level" for failure is set that is "as low as technologically feasible",[9] then, accidents will happen for which no one will be culpable (there will be a culpability gap), but this will not be a moral problem, because all relevant considerations in terms of safety, justice, and responsibility will be respected.

We agree that culpability should not be pursued always and at all costs, and that some culpability gaps are unavoidable and morally acceptable. However, we doubt that Simpson and Müller have identified a fair system to decide which culpability gaps with AI systems are morally acceptable. Indeed, justice and rights can be preserved by setting reasonable standards of care, which may also allow for fair culpability attribution and the prevention of unwanted responsibility gaps. But this is a statement of the problem, not its solution: it is close to question-begging. To what extent it is possible to establish standards of reasonable care for the design, use, and regulation of AI in the same way in which we do for buildings and bridges is precisely the question raised in present-day (legal) debate on the responsibility gaps for AI. In relation to US law, Ryan Calo pointed out that culpability gaps with AI may happen precisely because the traditional assumptions about what should count as sufficient intention, knowledge, and foreseeability on

[9] In their words: "If it is possible to build robots that comply with a strict tolerance level, then that is obligatory. But if it is not possible, the best we can do is good enough" (316–7).

the side of the defendant (criminal law) may not apply, due to the emergent and unpredictable behaviour of AI (Calo, 2015, p. 542). Also, traditional assumptions on designers or managers having "exclusive control" on artefacts can hardly apply in relation to systems whose behaviour is influenced by a chain of different actors (manufacturers, programmers, users). How to achieve a fair and right-respecting distribution of risk and culpability in this new context is still open to discussion. Moreover, the *accountability* gap potentially created by the unpredictable, opaque, and interactive behaviour of artificial intelligence may also make it difficult to establish a priori whether the system will be able to comply with the requirement of fairness: designers and users of artificial intelligence may simply not be in the position to know that a system is discriminatory or otherwise unfair in the first place.

A related concern with Simpson and Müller's proposal is that by stating that the "threshold of safety" eventually determining the culpability attribution should be "as low as technologically feasible"; they risk to encourage the belief that this is mainly a technical question, one that is up to engineers and "experts" to solve. However, giving engineers and other "experts" the power to decide on this threshold may result in a technocratic approach sharpening rather than solving the culpability gap: technical experts may (honestly) believe that nobody is to blame for an accident because they have done what could reasonably be expected from them, but this may not match with the well-informed judgement of a non-expert and the moral and legal requirements of society. In fact, recent legal history of the standards of negligence in medical practice shows a trend towards abandoning a system in which courts rely exclusively on expert professional opinion for their assessment of professional malpractice. This is due to the recognition that professional opinion may sometimes be "unreasonable" or "irresponsible", too conservative, biased, or otherwise reflecting the interests of the members of the profession rather than the interest of the public. Also from a normative point of view, courts being "dictated to" by experts were considered as a dangerous shift towards technocracy and not in line with a "right-based" society (Mulheron, 2010).

Relatedly, from a broader perspective, old-style division of labour – engineers give facts to regulators and regulators establish whether the technology is safe enough – does not incentivise mechanisms of moral accountability between engineers and societal stakeholders (Funtowicz & Ravetz, 1990), insofar as they may not sufficiently promote a well-informed public deliberation on what a "reasonable threshold" of safety (and other values) in emerging technology should be. Nor does this approach incentivise engineers to go beyond the current state of art in technology and look for innovative solutions which may improve the capabilities of current technology to better satisfy complex and potentially conflicting societal demands (Van den Hoven et al., 2012) such as, for instance, higher levels of safety combined with better predictability etc. In the terminology introduced above, deflationist strategies like Simpson and Müller's fail to address gaps in the active responsibility of engineers. They may also fail to address gaps in (public) accountability, insofar as they seem to delegate to experts the setting of a reasonable standard of care.

3.2 The Risks of Solutionism

Others recognise the novelty of the responsibility gap but do not believe in its intractability and have tried to offer some new technical and legal solutions to address it. Whereas some of these solutions might in principle be part of a comprehensive strategy to address the problem of responsibility gap in its entirety and complexity, their authors often fall short of providing such a comprehensive plan. Moreover, even when this is not the intention of their authors, these proposals even run the risk of fuelling the temptation of "solutionism" (Morozov, 2013); (Stilgoe, 2017), the belief that complex socio-technical and political problems can be "solved" (or avoided) by the introduction of new techniques. We distinguish here two main approaches to address the responsibility gaps, the technical and the legal, and we suggest that if taken in abstraction from the broader picture presented above, these run the risk of leading, correspondingly, to "technical solutionism" and "legal solutionism".

3.2.1 Explainable AI and "Technical Solutionism"

One of the commonly recognised causes of generically defined "responsibility gaps" is, as we have seen in the previous sections, the lack of transparency, explainability, and interpretability of machine-aided decision-making, be it defined as just algorithmic or properly AI. Although there being multiple senses and extents to which a system can be said to be explainable (Doran et al., 2017), we will stick to the most basic form, where "a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs", and where transparency of the whole process is granted. Theorists identified transparency and explainability of algorithms and AI as an important element to safeguard the "traceability" of human responsible agents and, consequently, a fair attribution of moral responsibility (Mittelstadt et al., 2016). We believe that algorithmic explainability, though constituting one interesting element in a complex strategy to address the responsibility gaps, is neither a sufficient nor a necessary condition to address them. Believing the contrary would amount to what we have called "technical solutionism": the belief that new technological solutions may be sufficient in themselves to address complex socio-technical and political problems.

One reason for explainability not being sufficient to address the responsibility gaps is that, as seen, such gaps are due to different problems that are not entirely addressed by increasing transparency and explainability of the algorithmic parts of a system. Mittelstadt et al. (2016, p. 12), echoing Simon (2015), correctly pointed out how an important factor determining traceability, still under-researched, is the fact that responsibility is distributed, or diffused, "across a network of human and algorithmic actors simultaneously", which is closely related to what we previously refer to as "the problem of many hands" and, more generally, "organisational black boxes". Relatedly, algorithmic transparency and explainability do not necessarily allow for fair attribution of moral and legal culpability. Very complex systems might be in principle understandable and open to scrutiny, but perhaps only "after the fact", by a very selected audience, and in a relatively large timeframe. But this does not mean that this behaviour may be sufficiently understood and predicted in advance by any of the human agents involved in their design use or regulation. Nor does this entail that any of these agents has been given a sufficient awareness and capacity to comply with some specific obligations, to prevent some outcome, to explain it once it has happened, or both.

Relatedly, explainability might in other cases also not be necessary to address culpability, accountability, and active responsibility gaps. Some culpability and accountability gaps with AI can be potentially addressed by providing some agents along the chain of design, development regulation, use with a sufficient knowledge of the limitations of the technical systems (including their opacity etc.), and a sufficient awareness of their obligation to prevent unwanted results in the deployment of a such technologies (Santoni de Sio & van den Hoven, 2018). That is by promoting their active responsibility. In the presence of sufficient knowledge and training, then, for instance, a military commander can be reasonably held accountable and culpable for his conscious decision to deploy an unpredictable technical system in a military mission, which ends up in the unlawful killing of innocent civilians. Similarly, the manager of a car manufacturing company and/or the chair of a road safety agency can be legitimately held accountable and culpable for their decision to put/allow on the public road a vehicle whose behaviour, as they well knew, could not be sufficiently predicted and explained. In a relevant sense, they could and should have known better.[10]

3.2.2 New Liability Regimes and the Risks of "Legal Solutionism"

Some legal scholars and policy-makers have recently recognised that the introduction of AI systems may potentially increase the number of accidents and/or introduce new kinds of accidents and/or increase the number of accidents for which victims may not receive compensation, due to the difficulty of applying existing legal liability regimes, typically negligence and product liability, to any of the actors involved in the network of design and use of new technologies: the legal culpability gap (Calo, 2015) (Pagallo, 2013).[11] To address these issues, they have committed themselves to work towards revising or introducing new liability mechanisms, which may allow for compensating victims of accidents involving AI for which no clear human fault can be attributed. Some of these regimes would be the faultless compensation schemes for damages caused by AI systems discussed by Schellekens (2018) and the introduction of electronic personhood proposed among others by the European Parliament resolution on Civil Law Rules on Robotics (Delvaux, 2017) and discussed among others by Koops et al. (2010) and, in a critical fashion, Bryson et al. (2017). However, an exclusive focus on bridging the liability gap may be insufficient and potentially self-defeating from the point of view of the broader plan of bridging the responsibility gaps.

[10] More controversial would be the application of this logic to the case of technology developers, due to their smaller power in the managerial decisions.

[11] See Sect. 2.1 above.

As we have explained in Section 2.1, there are several reasons why we might want to preserve fair practices of attribution of moral blame and culpability as well as “active responsibility” practices in addition to fair and effective practice of legal liability and compensation schemes. Answering the questions “who should be legally punished” or “who pays” (Pagallo 2015) for wrong AI-mediated decisions and behaviours is not sufficient to answer the broader question “who is responsible” for them and how to prevent these outcomes in the first place. Liability regimes grounded in individual culpability or fault (such as criminal liability and criminal and civil liability by negligence) might be well-suited to deal with clear and bold individual responsibilities. However, they might be less adequate to cope with substantial shared responsibilities derived from manifold individual small faults. According to an example of van de Poel et al. (2012), it would not probably make sense to hold individual people liable for their share of pollution, but that does not mean that they cannot be blamed or shamed for that, or that other policy and legislative tools should be used to discourage individual and corporate behaviour from increasing pollution. The same can be said about the effects of digital technologies.

Faultless liability regimes and legal personhood of artificial agents not only risk to shift away attention from culpability and accountability but also from active responsibility. In fact, those approaches might underestimate the importance of promoting proactive, preventive approaches to create safe and societally beneficial technical systems. Liability regimes are managed by the State and require strict standards of causation, evidence, and seriousness. Effective as they may be in (dis)incentivising some behaviour, those regimes do not and cannot cover all undesirable behaviour. It is not possible or desirable, e.g. to have the State checking and judging professional behaviour, but that does not mean that anything goes. Since risky behaviours without (provable) harm would fail to be sanctioned under a liability scheme, professionals’ good conduct can only be granted by relying on their own awareness and knowledge of their moral and legal responsibility towards society, and their individual capacity and motivation to comply with it. Moreover, corporate or civil liability may not be a strong incentive to behave, e.g. for agents or companies who can easily afford to pay any fine or compensation or may rely on the difficulties to enforce legal norms in this field[12]; blaming and shaming and strong political initiatives may sometimes be a more effective tool.

In addition to adjust and revise regimes of liability, we also need to create better mechanisms to promote the moral accountability of all agents involved in the design and use of AI systems; better mechanisms of public accountability for those who design or regulate AI systems operating in the public space; and, possibly more importantly, mechanisms and policies to promote a better culture of active responsibility of all the designers, managers, controllers, and users of AI systems.

[12] A prominent example being the recent story of the GDPR, that some critics summarised in: the “world’s toughest privacy law” proven “toothless” (Vinocur, 2019).

4 The Need of a Comprehensive Approach to Address the Responsibility Gaps and the Promises of "Meaningful Human Control"

Our analysis has shown that the problem of responsibility gap is not one problem but a set of at least four interconnected problems – gaps in culpability, moral and public accountability, active responsibility – caused by different sources, some technical, other organisational, legal, ethical, and societal. And that partial approaches to the responsibility gap – i.e. focusing only on one form of responsibility and/or only one source of gaps – not only bring the risk of offering an incomplete picture of the problem but also a distorted one, one that hinders the creation of an appropriate response (Fig. 1). In this last section, we will try to outline what a more comprehensive approach may look like (Table 2). We will do so by referring to what we consider to be a very promising approach to be found in the literature on ethics of AI: the recent interpretation of "meaningful human control" by Santoni de Sio and van den Hoven (2018). Future work will have to further develop and substantiate this proposal.

4.1 Meaningful Human Control: the Concept

Experts from different disciplines have recently converged towards the idea that granting "meaningful human control" over AI would substantially contribute to address responsibility gaps. Multiple accounts of meaningful human control (MHC henceforth) have been therefore proposed (Ekelhof, 2019). They mostly consist of sets of normative requirements to promote a legally, ethically, and societally acceptable form of human control. Originally proposed in the context of lethal autonomous weapon systems (Amoroso & Tamburrini, 2019; Article36, 2014; Chengeta, 2016; Ekelhof, 2019; Horowitz & Scharre, 2015; Moyes, 2016; Schwarz, 2018), MHC has been recently investigated in the field of automated driving systems (Calvert et al., 2018, 2019, 2020; Heikoop et al., 2019; Mecacci & Santoni de Sio, 2020; Santoni de Sio & van den Hoven, 2018) and medical automation (Braun et al., 2020; Ficuciello et al., 2019). Relatedly, the importance as well as the difficulty of defining the kind of "meaningful human involvement" required to avoid responsibility gaps in automated decision-making is highlighted in the art. 22(1) GDPR as interpreted by the Article 29 Working Party/EDPB.

In their seminal work, Santoni de Sio and Van Den Hoven (2018), produced a philosophical account of MHC that aims to: (a) a more solid theoretical framework, grounded in philosophical theory of moral responsibility and control (Fischer & Ravizza, 1998) and (b) a strongly design-oriented perspective, according to the value-sensitive design approach. Their ambition is to provide a unified conceptual framework which also provides some principles to practically (re) configure AI to minimise possible responsibility gaps, by acting at the level of technical as well as organisational and legal design. According to their theory, in order for AI systems to be under meaningful human control, two major

Table 2 Meaningful human control (MHC) promoting the four types of responsibility

Meaningful human control		
Conditions	Operationalisation	Payoffs for different types of responsibility
Tracking: alignment between system and relevant reasons of relevant human agents	Mapping the agents involved in the system, their reasons/intentions, and their relations to the systems	*Culpability and accountability* Extends culpability and accountability attribution to a wider range of agents (e.g. companies and regulators for self-driving cars accidents)
		Active responsibility Supports managers, designers, policy-makers, researchers in identifying potential value tensions and recognising their own responsibility to design for all the relevant reasons
Tracing: alignment between system and human capacities (technical, motivational and moral)	Analysing the capacities of the human agents in the system and checking that the system sufficiently reflects them	*Culpability* Allows for fairer attributions: avoiding both scapegoating and impunity for accidents (blaming the driver or nobody for an avoidable self-driving car accident)
		Moral accountability Requires that the relevant agents understand the AI working processes and their role in the decision-making process (e.g. the medical doctor using AI for diagnosis)
		Public accountability Requires that public officers (not only IT experts and private companies) remain able and motivated to understand the AI working processes and their duty to respond to the public for it
		Active responsibility Supports various actors in developing the knowledge, capacity, opportunity and motivation to discharge their different responsibilities in relation to AI

conditions have to be satisfied, called "tracking" and "tracing". These conditions describe, respectively, the nature of the control relation and the features that a human–machine system should strive for to maintain human responsibility on the system. Tracking requires that the socio-technical system – i.e. the whole combination of technical, human, and organisational elements – is designed to demonstrably respond to the relevant reasons of the relevant agents and to the relevant facts in the environment.[13] Tracing requires that the socio-technical system is designed so that for each of the (relevant) actions of the system, it is possible to identify at least one human agent along the chain of design, development, and use that possesses both (i) sufficient knowledge of the capabilities and limitations of the system and (ii) sufficient moral awareness of, and capacity to comply with, her role as potential target of legitimate response for the behaviour of the system. Tracking requires the alignment of the system with the values, reasons, and intentions of the relevant human agents; tracing requires the alignment of the system with the capacities of the relevant human agents. According to an *operational*, causal notion of control – mainly adopted in scientific and technical domains – a technical system is under the control of a human agent when there is a reliable causal connection between their and the machine's *behaviour*. The philosophical and normative idea behind MHC is that morally relevant control and moral responsibility depend on the socio-technical system being aligned with the *reasons* and the *capacities* of the relevant human agents (reasons and capacities not behaviour are the source of "meaningfulness").

4.2 The "Tracking" Condition and its Payoffs for Responsibility

In contrast with "technical solutionism", and in line with legal-oriented solutions, the MHC approach recognises that human control requires not only new technical features to promote technical transparency (explainable AI) but also different social, institutional, and legal arrangements to promote organisational and political transparency about the control roles. The "tracking condition" for MHC demands a systematic specification of the control roles based on the system "responsiveness" to human reasons. This requires as a first step a map of all the relevant agents involved in the design, control, regulation, and use of a system, as well as an analysis of the extent to which their reasons, values, and intentions are (or should be) reflected in its behaviour. Mecacci and Santoni de Sio (2020) have recently started developing such a map in relation to automated driving systems (dual-mode vehicles).

[13] An important clarification: the tracking condition does not necessarily require that *machines* respond to abstract values or that abstract reasons and values are hard-coded in the machine; it does require that the socio-technical system (human + machine) is responsive: the human elements can do the job. At the same time, also some "*autonomous systems*" (operating without human supervision) may remain under MHC under some circumstances, to the extent they are sufficiently capable to interact with the (morally) relevant aspects of the environment in which operate, e.g. a self-driving car operating on a dedicated lane with "readable" signs and relatively predictable interactions. Correspondingly, many autonomous systems may *not* be compatible with MHC, typically those operating in environments presenting relevant (moral) features too difficult to interpret for the machine (e.g. present-day battlefields), or learning systems able to reprogram themselves in ways that may not demonstrably track the reasons of designers, regulators, relevant users, etc. (including science-fictional artificial general intelligence).

As a further illustration, consider the recent Uber accident (Bellon, 2018; Stilgoe, 2020), in which a car in automated mode but with a human operator onboard ran over a pedestrian due to a technical failure in the sensing system. Who was in control of the system and thus responsible for the accident? The legal liability process clearly relied on a narrow notion of control: the onboard operator was deemed in control of the system as the vehicle was responsive to her behaviour (she was in the position to intervene and avoid the crash), and since she failed to behave as requested by her role, she was considered negligent *and* liable for the death (BBC News, 2020). This answer sounds too simplistic: more "controllers" are involved in this story. By using Mecacci and Santoni de Sio's (2020) approach, we can more easily identify other actors whose higher-level plans and goals played or could have played a control role in the story, and may therefore have some form of responsibility in it. Just to mention two: the vehicle's dangerous behaviour was clearly dependant also on the explicit intention of the owning company (Uber) to test new not-yet-safe self-driving features on the public road with the intention of pushing their technical development, and the opportunity to do these tests was created by the Arizona's authorities with the goal of attracting (big) tech companies in their State (National Transportation Safety Board, 2019). So it seems fair to attribute some *prima facie* moral culpability also to them for the death. This in turn triggers a moral duty for them to explain to the victim's relatives and to the broader public the reasons of their strategic and political choices (moral and public accountability). Correspondingly, the map offers a starting point for designers, policy-makers, and researchers to identify the different active responsibilities of the agents involved, as well as a general schema to identify the policy, legal, technical, and organisational interventions needed to ensure that the system is sufficiently responsive to the right reasons of the relevant agents.

4.3 The Tracing Conditions and its Payoffs for Responsibility

Unlike proposals based on new forms of legal liability, MHC proposes that socio-technical systems are also systematically designed to avoid gaps in moral culpability, accountability, and active responsibility. The "tracing condition" proposes that a system can remain under MHC only in the presence of a solid alignment between the system and the technical, motivational, moral capacities of the relevant agents involved, with different roles, in the design, control, and use of the system. The direct goal of this condition is promoting a fair distribution of moral culpability, thereby avoiding two undesired results: first, scapegoating, i.e. agents being held culpable without having a fair capacity to avoid wrongdoing (Elish, 2019): in the example of the automated driving systems above, for instance, the drivers' relevant technical and motivational capacities not being sufficiently studied and trained. Second, impunity for avoidable accidents, i.e. culpability gaps: the impossibility to legitimately blame anybody as no individual agent possesses all the relevant capacities, e.g. the managers/designers having the technical capacity but not the moral motivation to avoid accidents and the drivers having the motivation but not the skills. The tracing condition also helps addressing accountability and active responsibility gaps. If a person or organisation should be morally or publicly accountable, then they must also possess the specific capacity to discharge this duty: according to another example discussed above, if a doctor has to remain accountable to their patients for her decisions, then she should maintain the capacity and motivation to understand the functioning of the AI system she uses and to explain her decision to the patients. If a public organisation has to remain accountable to their citizens, then arguably public officers and civil servants, not IT experts and private companies, should maintain the capacity and motivation to understand and explain the behaviour of their institution to the citizens, no matter how much technology is involved in the decision. So, the tracing condition poses a strong accent on the individual, context-relative capacities needed for the relevant human actors to remain accountable to different subjects and fora and in relation to different activities. The analysis and promotion of (moral) capacities are also key to ensure the discharge of various active responsibilities by various actors. The responsible innovation approach invites stakeholders to be responsive to one another and to the relevant values in the process of technological development (Stilgoe et al., 2013). The tracing condition also demands that these stakeholders are effectively supported in acquiring the context-specific capacity, opportunity, and motivation to do so. For instance, across the years, the aviation sector has built up a culture of safety or the medical profession a culture of care: as also recently proposed by the Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility, (Bonnefon et al., 2020), a new "culture of responsibility" – new practices, incentives, identities – should be promoted within private and public organisations in relation to the design and use of AI-based systems.

5 Conclusion

To improve the understanding of the problem of the "responsibility gap" for artificial intelligence (AI), we have proposed to rely on a comprehensive analysis of four forms of responsibility presented in some relevant philosophical and legal literature. A first result of this analysis is a reasoned map of four potential responsibility gaps in socio-technical systems that include AI (Table 1). A second result is the presentation of some limits of the current, partial, approaches to responsibility gaps, which fail to identify the complexity of the responsibility gap problems (Fig. 1). A third result is the sketch of a proposal for a more integrated and comprehensive approach to address, by design, the four responsibility gaps in their interrelations. The proposal is grounded in one particular interpretation of "meaningful human control" (Table 2). Based on these results, future philosophical, empirical, and technical work will have, among other things, to further clarify which responsibility gaps (may) emerge in relation to different systems in different contexts of application, and the extent to which the approach to meaningful human control here presented may effectively address these gaps.

Acknowledgements The authors are grateful to the Meaningful Human Control over Automated Driving Systems team for helping to shape some ideas contained in this paper. The content of this paper has also been discussed within the Horizon 2020 Commission group on ethical issues raised by driverless mobility (E03659), of which Filippo Santoni de Sio was member. Filippo Santoni de Sio is grateful to the Expert Group, and in particular to John Danaher and Karolina Zawieska, for contributing to develop some of the ideas of this paper. Filippo Santoni de Sio is also grateful to Mark Bovens for helping to clarify the nature of public accountability (gaps) in relation to AI, in a series of private conversations, and to Herman Veluwenkamp for some written feedback on a previous version of this paper. Giulio Mecacci is grateful to Djamari Oetringer for her input on deep learning techniques. Finally, the authors are very grateful to two anonymous reviewers for their helpful comments and constructive criticisms on a previous version. The authors remain responsible for the final text, in all senses.

Authors' Contributions Filippo Santoni de Sio outlined the main concept of the paper and lead the writing process. Both authors have contributed to develop the ideas presented in the paper and have drafted and revised the text multiple times.

Funding The paper is part of the research project Meaningful Human Control over Automated Driving System (MHC-ADS) financed by the Netherlands Organisation for Scientific Research (NWO) with project number MVI.16.044.

Data Availability N.A.

Code Availability N.A.

Declarations

Conflicts of Interest N.A.

References

Amoroso, D., & Tamburrini, G. (2019). What makes human control over weapon systems "meaningful"? *ICRAC Working Paper Series #4*.

Article 36. (2014). *Autonomous weapons, meaningful human control and the CCW*.

Asaro, P. (2012). A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics, MIT Press* (pp. 169–186).

BBC News. (2020). *Uber's self-driving operator charged over fatal crash*. https://www.bbc.com/news/technology-54175359. Accessed 8 April 2021

Bellon, T. (2018). *Fatal U.S. self-driving auto accident raises novel legal questions*. Reuters. https://www.reuters.com/article/us-autos-selfdriving-uber-liability-anal/fatal-u-s-self-driving-auto-accident-raises-novel-legal-questions-idUSKBN1GW2SP. Accessed 8 April 2021

Bonnefon, J.-F., Černy, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., Martens, M., Mladenovič, M., Palade, P., Reed, N., Santoni de Sio, F., Tsinorema, S., Wachter, S., & Zawieska, K. (2020). *Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*. Publication Office of the European Union: Luxembourg.

Bovens, M. (1998). *The quest for responsibility: Accountability and citizenship in complex organisations*. Cambridge University Press.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal, 13*(4), 447–468.

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review, 62*(2), 174–184. https://doi.org/10.1111/0033-3352.00168.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics*, medethics-2019–105860. https://doi.org/10.1136/medethics-2019-105860

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law, 25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9.

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*(3), 513–563. https://doi.org/10.2139/ssrn.2402972.

Calvert, S. C., Mecacci, G., Heikoop, D. D., & Santoni de Sio, F. (2018). Full platoon control in Truck platooning: A meaningful human control perspective. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3320–3326. https://doi.org/10.1109/ITSC.2018.8570013

Calvert, S. C., Heikoop, D. D., Mecacci, G., & van Arem, B. (2019). A human centric framework for the analysis of automated driving systems based on meaningful human control. *Theoretical Issues in Ergonomics Science*, *0*(0), 1–29.https://doi.org/10.1080/1463922X.2019.1697390

Calvert, S. C., Mecacci, G., van Arem, B., Santoni de Sio, F., Heikoop, D. D., & Hagenzieker, M. (2020). Gaps in the Control of automated vehicles on roads. *IEEE Intelligent Transportation Systems Magazine*, 1–1.https://doi.org/10.1109/MITS.2019.2926278

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20–23. https://doi.org/10.1038/538020a.

Chengeta, T. (2016). Defining the emerging notion of meaningful human controll in autonomous weapon systems (AWS). *NYU Journal of International Law*. https://doi.org/10.2139/ssrn.2754995.

Coeckelbergh, M. (2019). Artificial Intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *7*(0123456789). https://doi.org/10.1007/s11948-019-00146-8

Collingridge, D. (1980). The Social Control of. *Technology*. https://doi.org/10.2307/2634327.

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

Delvaux, M. (2017). *Report with recommendations to the Commission on Civil Law Rules on Robotics (A8–0005/2017).*

Di Nucci, E., & Santoni de Sio, F. (2016). *Drones and responsibility: mapping the field*. Routledge.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable ai really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a right to explanationn is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*(1), 1–65. https://doi.org/10.2139/ssrn.2972855.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348. https://doi.org/10.1111/1758-5899.12665.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*(0), 40. https://doi.org/10.17351/ests2019.260

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Martin's Press.

European Commission. (2019). *High-level expert group on artificial intelligence ETHICS GUIDELINES FOR TRUSTWORTHY AI.*

Feinberg, J. (1965). The Expressive function of punishment. *The Monist, 49*(3), 397–423. https://doi.org/10.5840/monist196549326.

Feinberg, J. (1970). *Doing & deserving; essays in the theory of responsibility*. Princeton University Press.

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., & Siciliano, B. (2019). Autonomy in surgical robots and its meaningful human control. *Paladyn, Journal of Behavioral Robotics, 10*(1), 30–43. https://doi.org/10.1515/pjbr-2019-0002.

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

Funtowicz, S. O., & Ravetz, J. R. (1990). Post-normal science: A new science for new times, October 1990, *Scientific European*, 20–22.

Gardner, J. (2007). The mark of responsibility. In *Offences and Defences* (pp. 177–200). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199239351.003.0009

Genus, A., & Stirling, A. (2018). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research Policy*. https://doi.org/10.1016/j.respol.2017.09.012.

Hart, H. L. A. (1968). *Punishment and responsibility*. Oxford University Press.

Hayenhjelm, M., & Wolff, J. (2012). The Moral Problem of Risk Impositions: A survey of the literature. *European Journal of Philosophy, 20*, E26–E51. https://doi.org/10.1111/j.1468-0378.2011.00482.x.

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., & van Arem, B. (2019). Human behaviour with automated driving systems: Aquantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science, 20*(6), 711–730. https://doi.org/10.1080/1463922X.2019.1574931.

Heyns, C. (2013). *Report of the Special Rapporteur on Extra-Judicial, Summary or Arbitrary Executions,* United Nations.

Hildebrandt, M. (2019). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. In *Theoretical Inquiries in Law* 20(1).

Honoré, T. (1999). *Responsibility and Fault*. Hart Publishing.

Horowitz, M. C., & Scharre, P. (2015). *Meaningful human control in weapons systems: A primer*. Center for a New American Security.

Human Right Watch. (2015). *Mind the gap: The lack of accountability for killer robots*.

Koops, B.-J., Hildebrandt, M., & Jaquet-Chiffelle, D.-O. (2010). Bridging the accountability gap: Rights for new entities in the information society. *Minnesota Journal of Law, Science and Technology*.

Lindblom, C. E. (1990). *Inquiry and change : The troubled attempt to understand and shape society*. Yale University Press.

Lipsky, M. (1980). *Street-level bureaucracy : Dilemmas of the individual in public services*. Russell Sage Foundation.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1.

McKenna, M. (2012). Conversation and responsibility. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780199740031.001.0001.

Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology, 22*, 103–115. https://doi.org/10.1007/s10676-019-09519-w.

Meloni, C. (2016). State and individual responsibility for targeted killings by drones. In E. Di Nucci & F. Santoni de Sio (Eds.), *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Re-motely Controlled Weapons*. Routledge.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679.

Morozov, E. (2013). To save everything, click here: the folly of technological solutionism. In *To save everything, click here : the folly of technological solutionism*. Public Affairs.

Morse, S. J. (2006). Moral and legal responsibility and the new neuroscience. In J. Illes (Ed.), *Neuroethics in the 21st Century: Defining the Issues in Theory, Practice and Policy 33*. Oxford University Press.

Moyes, R. (2016). *Key Elements of Meaningful Human Control*. Article 36.

Mulheron, R. (2010). Trumping Bolam: A Critical Legal Analysis of Bolitho's "Gloss". *The Cambridge Law Journal, 69*(3), 609–638. https://doi.org/10.1017/S0008197310000826.

National Transportation Safety Board. (2019). *Collision Between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona March 18, 2018*.

Noto La Diega, G. (2018). Against the dehumanisation of decision-making – Algorithmic decisions at the crossroads of intellectual property, data protection, and freedom of information. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 19*(1).

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x.

Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts*. Springer.

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061.

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pereboom, D. (2006). *Living without free will*. Cambridge University Press.

Pesch, U. (2015). Engineers and Active Responsibility. *Science and Engineering Ethics, 21*(4), 925–939. https://doi.org/10.1007/s11948-014-9571-7.

Santoni de Sio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice*, 411–429.https://doi.org/10.1007/s10677-017-9780-7

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2018.00015.

Schellekens, M. (2018). No-fault compensation schemes for self-driving vehicles. *Law, Innovation and Technology, 10*(2), 314–333. https://doi.org/10.1080/17579961.2018.1527477.

Schwarz, E. (2018). *The (im)possibility of Meaningful human control for lethal autonomous weapon systems*. https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/.

Sie, M. (2005). *Justifying blame: Why free will matters and why it does not*. Rodopi.

Simon, J. (2015). Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (Ed.), *The Onlife Manifesto* (pp. 145–159). Springer International Publishing. https://doi.org/10.1007/978-3-319-04093-6_17

Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *The Philosophical Quarterly, 66*(263), 302–322. https://doi.org/10.1093/pq/pqv075.

Sini, C. (2021). Machine, culture, and robot. In S. Chiodo & V. Schiaffonati (Eds.), *Italian Philosophy of Technology* (pp. 83–88). Springer.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Stilgoe, J. (2017). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, 030631271774168. https://doi.org/10.1177/0306312717741687

Stilgoe, J. (2020). Who Killed Elaine Herzberg? In *Who's Driving Innovation?* (pp. 1–6). Springer International Publishing. https://doi.org/10.1007/978-3-030-32320-2_1

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy, 42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008.

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

Swierstra, T., & Jelsma, J. (2006). Responsibility without Moralism in Technoscientific Design Practice. *Science, Technology, & Human Values, 31*(3), 309–332. https://doi.org/10.1177/0162243905285844.

Thompson, D. F. (1980). Moral responsibility of public officials : The problem of many hands. *The American Political Science Review, 74*(4), 905–916.

Van de Poel, I., & Sand, M. (2018). Varieties of responsibility: two problems of responsible innovation. *Synthese*. https://doi.org/10.1007/s11229-018-01951-7.

Van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The Problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0.

Van de Poel, I., Royakkers, L. M. M., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. . Routledge. https://doi.org/10.4324/9781315734217.

Van den Hoven, J., Lokhorst, G. J., & Van de Poel, I. (2012). Engineering and the Problem of Moral Overload. *Science and Engineering Ethics, 18*(1), 143–155. https://doi.org/10.1007/s11948-011-9277-z.

Van Eck, M. (2018). Geautomatiseerde ketenbesluiten & rechtsbescherming: Een onderzoek naar de praktijk van geautomatiseerde ketenbesluiten over een financieel belang in relatie tot rechtsbescherming. PhD Dissertation, Tilburg University. https://research.tilburguniversity.edu/en/publications/automated-administrative-chain-decisions-amp-legal-protection-res.

Vinocur, N. (2019). 'We have a huge problem': European tech regulator despairs over lack of enforcement. *Politico*. https://www.politico.eu/article/we-have-a-huge-problem-european-regulator-despairs-over-lack-of-enforcement/. Accessed 8 April 2021.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005.

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.

Zouridis, S., Bovens, M., & Van Eck, M. (2019). Digital discretion. In T. Evans & P. Hupe (Eds.), *Discretion and the quest for controlled freedom*. Palgrave/MacMillan.