PDF version of the entry
Computing and Moral Responsibility
https://plato.stanford.edu/archives/spr2023/entries/computing-responsibility/

from the Spring 2023 Edition of the

Stanford Encyclopedia
of Philosophy

Library of Congress ISSN: 1095-5054

Computing and Moral Responsibility

First published Wed Jul 18, 2012; substantive revision Thu Feb 2, 2023

Traditionally philosophical discussions on moral responsibility have focused on the human components of moral action. Accounts of how to ascribe moral responsibility usually describe human agents performing actions that have well-defined, direct consequences. In today's increasingly technological society, however, human activity cannot be properly understood without making reference to technological artifacts, which complicates the ascription of moral responsibility (Jonas 1984; Doorn & van de Poel 2012).[1] As we interact with and through these artifacts, they affect the decisions that we make and how we make them (Latour 1992, Verbeek 2021). They persuade, facilitate and enable particular human cognitive processes, actions or attitudes, while constraining, discouraging and inhibiting others. For instance, internet search engines prioritize and present information in a particular order, thereby influencing what internet users get to see. As Verbeek points out, such technological artifacts are "active mediators" that "actively co-shape people's being in the world: their perception and actions, experience and existence" (2006, p. 364). As active mediators, they are a key part of human action and as a result they challenge conventional notions of moral responsibility that do not account for the active role of technology (Jonas 1984; Johnson 2001; Swierstra and Waelbers 2012).

Computing presents a particular case for understanding the role of technology in moral responsibility. As computer technologies have become a more integral part of daily activities, automate more decision-making processes and continue to transform the way people communicate and relate to each other, they have further complicated the already problematic tasks of attributing moral responsibility. The growing pervasiveness of computer technologies in everyday life, the growing

complexities of these technologies and the new possibilities that they provide raise new kinds of questions: who is responsible for the information published on the Internet? To what extent and for what period of time are developers of computer technologies accountable for untoward consequences of their products? And as computer technologies become more complex and behave increasingly autonomous can or should humans still be held responsible for the behavior of these technologies?

This entry will first look at the challenges that computing poses to conventional notions of moral responsibility. The discussion will then review two different ways in which various authors have addressed these challenges: 1) by reconsidering the idea of moral agency and 2) by rethinking the concept of moral responsibility itself.

---

1. Challenges to moral responsibility

Moral responsibility is about human action and its intentions and consequences (Fisher 1999, Eshleman 2016, Talbert 2022). Generally speaking a person or a group of people is morally responsible when their voluntary actions have morally significant outcomes that would make it appropriate to blame or praise them. Thus, we may consider it a person's moral responsibility to jump in the water and try to rescue another person, when she sees that person drowning. If she manages to pull the person from the water we are likely to praise her, whereas if she refuses to help we may blame her. Ascribing moral responsibility establishes a link between a person or a group of people and someone or something affected by the actions of this person or group. The person or group that performs the action and causes something to happen is often referred to as the *agent*. The person, group or thing that is affected by the action is referred to as the *patient*. Establishing a link in terms of moral responsibility between the agent and the patient can be done both retrospectively as well as prospectively. That is, sometimes ascriptions of responsibility involve giving an account of who was at fault for an accident and who should be punished. It can also be about prospectively determining the obligations and duties a person has to fulfill in the future and what she ought to do.

However, the circumstances under which it is appropriate to ascribe moral responsibility are not always clear. On the one hand the concept has varying meanings and debates continue on what sets moral responsibility apart from other kinds of responsibility (Hart 1968, Talbert 2022, Tigard 2021a). The concept is intertwined and sometimes overlaps with notions of accountability, liability, blameworthiness, role-responsibility and

causality. Opinions also differ on which conditions warrant the attribution of moral responsibility; whether it requires an agent with free will or not and whether humans are the only entities to which moral responsibility can be attributed (see the entry on moral responsibility).

On the other hand, it can be difficult to establish a direct link between the agent and the patient because of the complexity involved in human activity, in particular in today's technological society. Individuals and institutions generally act with and in *sociotechnical* systems in which tasks are distributed among human and technological components, which mutually affect each other in different ways depending on the context (Bijker, Hughes and Pinch 1987, Felt et al. 2016). Increasingly complex technologies can exacerbate the difficulty of identifying who or what is 'responsible'. When something goes wrong, a retrospective account of what happened is expected and the more complex the system, the more challenging is the task of ascribing responsibility (Johnson and Powers 2005). Indeed, Matthias argues that there is a growing 'responsibility gap': the more complex computer technologies become and the less human beings can directly control or intervene in the behavior of these technologies, the less we can reasonably hold human beings responsible for these technologies (Matthias, 2004).

The increasing pervasiveness of computer technologies poses various challenges to figuring out what moral responsibility entails and how it should be properly ascribed. To explain how computing complicates the ascription of responsibility we have to consider the conditions under which it makes sense to hold someone responsible. Despite the ongoing philosophical debates on the issue, most analysis of moral responsibility share at least the following three conditions (Eshleman 2016; Jonas 1984):

1. There should be a causal connection between the person and the outcome of actions. A person is usually only held responsible if they

had some control over the outcome of events.
2. The subject has to have knowledge of and be able to consider the possible consequences of her actions. We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.
3. The subject has to be able to freely choose to act in certain way. That is, it does not make sense to hold someone responsible for a harmful event if her actions were completely determined by outside forces.

A closer look at these three conditions shows that computing can complicate the applicability of each of these conditions.

1.1 Causal contribution

In order for a person to be held morally responsible for a particular event, she has to be able to exert some kind of influence on that event. It does not make sense to blame someone for an accident if she could not have avoided it by acting differently or if she had no control over the events leading up to the accident.

However, computer technologies can obscure the causal connections between a person's actions and the eventual consequences. Tracing the sequence of events that led to a computer-related catastrophic incident, such as a plane crash, usually leads in many directions, as such incidents are seldom the result of a single error or mishap. Technological accidents are commonly the product of an accumulation of mistakes, misunderstanding or negligent behavior of various individuals involved in the development, use and maintenance of computer systems, including designers, engineers, technicians, regulators, managers, users, manufacturers, sellers, resellers and even policy makers.

The involvement of multiple actors in the development and deployment of technologies gives rise to what is known as the problem of 'many hands': it is difficult to determine who was responsible for what when multiple individuals contributed to the outcome of events (Jonas 1984; Friedman 1990; Nissenbaum 1994; van de Poel et al. 2015). One classic example of the problem of many hands in computing is the case of the malfunctioning radiation treatment machine Therac-25 (Leveson and Turner 1993; Leveson 1995). This computer-controlled machine was designed for the radiation treatment of cancer patients as well as for X-rays. During a two-year period in the 1980s the machine massively overdosed six patients, contributing to the eventual death of three of them. These incidents were the result of the combination of a number of factors, including software errors, inadequate testing and quality assurance, exaggerated claims about the reliability, bad interface design, overconfidence in software design, and inadequate investigation or follow-up on accident reports. Nevertheless, in their analysis of the events Leveson and Turner conclude that it is hard to place the blame on a single person. The actions or negligence of all those involved might not have proven fatal were it not for the other contributing events. A more recent example of the problem of many hands is the crash of two 737 MAX passenger aircraft in late 2018 and early 2019. These crashes led to mulitple investigations that highlighted various factors that contributed to the tragic outcome, include design and human errors as well as organizational culture and lack of training (Heckert et al 2020). This is not to say that there is no moral responsibility in these case (Nissenbaum 1994; Gotterbarn 2001; Coeckelbergh 2012; Floridi 2013, Santonio De Sio et al, 2021), as many actors could have acted differently, but it makes it more difficult to retrospectively identify the appropriate person that can be called upon to answer and make amends for the outcome.

Adding to the problem of many hands is the temporal and physical distance that computing creates between a person and the consequences of her actions, as this distance can blur the causal connection between actions

and events (Friedman 1990). Computational technologies extend the reach of human activity through time and space. With the help of social media and communication technologies people can interact with others on the other side of the world. Satellites and advanced communication technologies allow pilots to fly a remotely controlled drone from their ground-control station half way across the world. These technologies enable people to act over greater distances, but this remoteness can dissociate the original actions from its eventual consequences (Waelbers 2009; Polder-Verkiel 2012; Coeckelbergh 2013). When a person uses a technological artifact to perform an action thousands of miles a way, that person might not know the people that will be affected and she might not directly, or only partially, experience the consequences. This can reduce the sense of responsibility the person feels and it may interfere with her ability to fully comprehend the significance of her actions. Similarly, the designers of an automated decision-making system determine ahead of time how decisions should be made, but they will rarely see how these decisions will impact the individuals they affect. Their original actions in programming the system may have effects on people years later.

The problem of many hands and the distancing effects of the use of technology illustrate the mediating role of technological artifacts in the confusion about moral responsibility. Technological artifacts bring together the various different intentions of their creators and users. People create and deploy technologies with the objective of producing some effect in the world. Software developers develop an automated content moderation tool, often at the request of their managers or clients, with the aim of shielding particular content from users and influencing what these users can or cannot read. The software has inscribed in its design the various intentions of the developers, managers and clients; it is poised to behave, given a particular input, according to their ideas about which information is appropriate (Friedman 1997, Gorwa, Binns, & Katzenbach 2020). Moral responsibility can therefore not be attributed without looking

at the causal efficacy of these artifacts and how they constrain and enable particular human activities.

However, although technological artefacts may influence and shape human action, they do not determine it. They are not isolated instruments that mean and work the same regardless of why, by whom, and in what context they are used; they have interpretive flexibility (Bijker et al. 1987) or multistability (Ihde 1990).[2] Although the design of the technology provides a set of conditions for action, the form and meaning of these actions is the result of how human agents choose to use these technologies in particular contexts. People often use technologies in ways unforeseen by their designers. This interpretive flexibility makes it difficult for designers to anticipate all the possible outcomes of the use of their technologies. The mediating role of computer technologies complicates the effort of retrospectively tracing back the causal connection between actions and outcomes, but it also complicates forward-looking responsibility.

1.2 Considering the consequences

As computer technologies shape how people perceive and experience the world, they affect the second condition for attributing moral responsibility. In order to make appropriate decisions a person has to be able to consider and deliberate about the consequences of their actions. They have to be aware of the possible risks or harms that their actions might cause. It is unfair to hold someone responsible for something if they could not have reasonably known that their actions might lead to harm.

On the one hand, computer technologies can help users to think through what their actions or choices may lead to. They help the user to capture, store, organize and analyze data and information (Zuboff 1982). For example, one often-named advantage of remote-controlled robots used by

the armed forces or rescue workers is that they enable their operators to acquire information that would not be able available without them. They allow their operators to look "beyond the next hill" or "around the next corner" and they can thus help operators to reflect on what the consequences of particular tactical decisions might be (US Department of Defense 2009). Similarly, data analysis tools can find patterns in large volumes of data that human data analysts cannot manually process (Boyd and Crawford 2012).

On the other hand the use of computers can constrain the ability of users to understand or consider the outcomes of their actions. These complex technologies, which are never fully free from errors, increasingly hide the automated processes behind the interface (Van den Hoven 2002). An example that illustrates how computer technologies can limit understanding of the outcomes are the controversial risk assessment tools used by judges in several states in the U.S. for parole decisions and sentencing. In 2016 a civil society organization found, based on an analysis of the risk scores of 7000 defendants produced by one particular algorithm, that the scores poorly reflected the actual recidivism rate and seemed to have a racial bias (Angwin et al. 2016). Regardless of whether its findings were correct or not, what is particularly relevant here is that the investigation also showed that judges did not have a full understanding of how the probabilities were calculated, in part because the algorithm was proprietary. The judges were basing their sentencing on the suggestion of an algorithm that they did not fully understand. This is the case for most computer technologies today. Users only see part of the many computations that a computer performs and are for the most part unaware of how it performs them; they usually only have a partial understanding of the assumptions, models and theories on which the information on their computer screen is based. The increasing complexity of computer systems and their reliance on opaque machine learning algorithms makes it even

more difficult to understand what is happening behind the interface (Pasquale 2015, Diakopoulos 2020).

The opacity of many computer systems can get in the way of assessing the validity and relevance of the information and can prevent a user from making appropriate decisions. People have a tendency to either rely too much or not enough on the accuracy automated systems (Cummings 2004; Parasuraman & Riley 1997). This tendency is called *automation bias*. A person's ability to act responsibly, for example, can suffer when she distrust the automation as result of a high rate of false alarms. In the Therac 25 case, one of the machine's operators testified that she had become used to the many cryptic error messages the machine gave and most did not involve patient safety (Leveson and Turner 1993, p.24). She tended to ignore them and therefore failed to notice when the machine was set to overdose a patient. Too much reliance on automated systems can have equally disastrous consequences. In 1988 the missile cruiser U.S.S. Vincennes shot down an Iranian civilian jet airliner, killing all 290 passengers onboard, after it mistakenly identified the airliner as an attacking military aircraft (Gray 1997). The cruiser was equipped with an Aegis defensive system that could automatically track and target incoming missiles and enemy aircrafts. Analyses of the events leading up to the incident showed that overconfidence in the abilities of the Aegis system prevented others from intervening when they could have. Two other warships nearby had correctly identified the aircraft as civilian. Yet, they did not dispute the Vincennes' identification of the aircraft as a military aircraft. In a later explanation Lt. Richard Thomas of one of the nearby ships stated, "We called her Robocruiser… she always seemed to have a picture… She always seemed to be telling everybody to get on or off the link as though her picture was better" (as quoted in Gray 1997, p. 34). The captains of both ships thought that the sophisticated Aegis system provided the crew of Vincennes with information they did not have.

Considering the possible consequences of one's actions is further complicated as computer technologies make it possible for humans to do things that they could not do before. Several decades ago, the philosopher Ladd pointed out, "[C]omputer technology has created new modes of conduct and new social institutions, new vices and new virtues, new ways of helping and new ways of abusing other people" (Ladd 1989, p. 210–11). Computer technologies of today have had a similar effect. The social or legal conventions that govern what we can do with these technologies take some time to emerge and the initial absence of these conventions contributes to confusion about responsibilities (Taddeo and Floridi 2015). For example, the ability for users to upload and share text, videos and images publicly on the Internet raised a whole set of questions about who is responsible for the content of the uploaded material. Such questions were at the heart of the debate about the conviction of three Google executives in Italy for a violation of the data protection act (Sartor and Viola de Azevedo Cunha 2010). The case concerned a video on YouTube of four students assaulting a disabled person. In response to a request by the Italian Postal Police, Google, as owner of YouTube, took the video down two months after the students uploaded it. The judge, nonetheless, ruled that Google was criminally liable for processing the video without taking adequate precautionary measures to avoid privacy violations. The judge also held Google liable for failing to adequately inform the students, who uploaded the videos, of their data protection obligations (p. 367). In the ensuing debate about the verdict, those critical of the ruling insisted that it threatened the freedom of expression on the Internet and it sets a dangerous precedent that can be used by authoritarian regimes to justify web censorship (see also Singel 2010). Moreover, they claimed that platform providers could not be held responsible for the actions of their users, as they could not realistically approve every upload and it was not their job to censure. Yet, others instead argued that it would be immoral for Google to be exempt from liability for the damage that others suffered

due to Google's profitable commercial activity. Cases like this one show that in the confusion about the possibilities and limitations of new technologies it can be difficult to determine one's moral obligations to others.

The lack of experience with new technological innovations can also affect what counts as negligent use of the technology. In order to operate a new computer system, users typically have to go through a process of training and familiarization with the system. It requires skill and experience to understand and imagine how the system will behave (Coeckelbergh and Wackers 2007). Friedman describes the case of a programmer who invented and was experimenting with a 'computer worm', a piece of code that can replicate itself. At the time this was a relatively new computational entity (1990). The programmer released the worm on the Internet, but the experiment quickly got out of the control when the code replicated much faster than he had expected (see also Denning 1989). Today we would not find this a satisfactory excuse, familiar as we have become with computer worms, viruses and other forms of malware. However, Friedman poses the question of whether the programmer really acted in a negligent way if the consequences were truly unanticipated. Does the computer community's lack of experience with a particular type of computational entity influence what we judge to be negligent behavior?

1.3 Free to act

The freedom to act is probably the most important condition for attributing moral responsibility and also one of the most contested (Talbert 2022). We tend to excuse people from moral blame if they had no other choice but to act in the way that they did. We typically do not hold people responsible if they were coerced or forced to take particular actions. In moral philosophy, the freedom to act can also mean that a person has free will or autonomy (Fisher 1999). Someone can be held morally responsible

because she acts on the basis of her own authentic thoughts and motivations and has the capacity to control her behavior (Johnson 2001). Note that this conception of autonomy is different from the way the term 'autonomy' is often used in computer science, where it tends to refer to the ability of a robot or computer system to independently perform (i.e. without the 'human in the loop') complex tasks in unpredictable environments for extended periods of time (Noorman 2009, Zerilli et al 2021).

Nevertheless, there is little consensus on what capacities human beings have, that other entities do not have, which enables them to act freely (see the entries on free will, autonomy in moral and political philosophy, personal autonomy and compatibilism). Does it require rationality, emotion, intentionality or cognition? Indeed, one important debate in moral philosophy centers on the question of whether human beings really have autonomy or free will? And, if not, can moral responsibility still be attributed (Talbert 2022)?

In practice, attributing autonomy or free will to humans on the basis of the fulfillment of a set of conditions turns out to be a less than straightforward endeavor. We attribute autonomy to persons in degrees. An adult is generally considered to be more autonomous than a child. As individuals in a society our autonomy is thought to vary because we are manipulated, controlled or influenced by forces outside of ourselves, such as by our parents or through peer pressure. Moreover, internal physical or psychological influences, such as addictions or mental problems, are perceived as further constraining the autonomy of a person.

Computing, like other technologies, adds an additional layer of complexity to determining whether someone is free to act, as it affects the choices that humans have and how they make them. One of the biggest application areas of computing is the automation of decision-making processes and

control. Automation can help to centralize and increase control over multiple processes for those in charge, while it limits the discretionary power of human operators on the lower-end of the decision-making chain. An example is provided by the automation of decision-making in public administration (Bovens and Zouridis 2002). Large public sector organizations have over the last decades progressively standardized and formalized their production processes. In a number of countries, the process of issuing decisions about (student) loans, social benefits, speeding tickets or tax returns is carried out to a significant extent by computer systems. This has reduced the scope of the administrative discretion that many officials, such as tax inspectors, welfare workers, and policy officers, have in deciding how to apply formal policy rules in individual cases (Eubanks 2018). In some cases, citizens no longer interact with officials that have significant responsibility in applying their knowledge of the rules and regulations to decide what is appropriate (e.g., would it be better to let someone off with a warning or is a speeding ticket required?). Rather, decisions are pre-programmed in the algorithms that apply the same measures and rules regardless of the person or the context (e.g., a speeding camera does not care about the context or personal circumstances), and the human beings that citizens do interact with have little opportunity to interrogate or change decisions (Dignum 2020). Responsibility for decisions made, in these cases, has moved from 'street-level bureaucrats' to the 'system-level bureaucrats', such as managers and computer experts, that decide on how to convert policy and legal frameworks into algorithms and decision-trees (Bovens and Zouridis 2002).

The automation of bureaucratic processes illustrates that some computer technologies are intentionally designed to limit the discretion of some human beings. An example is the anti-alcohol lock that is already in use in a number of countries, including the USA, Canada, Sweden and the UK. It requires the driver to pass a breathing test before she can start the car. This

technology forces a particular kind of action and leaves the driver with hardly any choice. Other technologies might have a more subtle way of steering behavior, by either persuading or nudging users (Verbeek 2016). For example, the onboard computer devices in some cars that show, in real-time, information about fuel consumption can encourage the driver to optimize fuel efficiency. Such technologies are designed with the explicit aim of making humans behave responsibly by limiting their options or persuading them to choose in a certain way.

Not all these technologies are designed to stimulate morally good behavior. Yeung notes that these kinds of decision-guidance techniques have become a key element of current day Big-Data analytic techniques, as used on social media and in advertising. She argues that these 'hyper nudges' are extremely powerful techniques to manipulate the behaviour of internet users and users of Internet of Things (IoT) devices due to their networked, continuously updated, dynamic and pervasive nature. As they gather data from a wide range of sources about users to continuously make predictions in real-time about the habits and preferences of users, they can target advertisement, information and price incentives to gently and unobtrusively nudge these users in directions preferred by the those that control the algorithms (Yeung, 2017). When these nudges are hardly noticeable and have a powerful effect, one can wonder how autonomous the decision making of these users is. This is the case, for example, with dark patterns, which use interfaces on websites or apps that are designed to trick users to do things that did not intend to do, such as purchasing additional expensive insurance (Ravenscraft 2020).

Verbeek notes that critics of the idea of intentionally developing technology to enforce morally desirable behavior have argued that it jettisons the democratic principles of our society and threatens human dignity. They argue that it deprives humans of their ability and rights to make deliberate decisions and to act voluntarily. In addition, critics have

claimed that if humans are not acting freely, their actions cannot be considered moral. These objections can be countered, as Verbeek argues, by pointing to the rules, norms, regulations and a host of technological artifacts that already set conditions for actions that humans are able or allowed to perform. Moreover, he notes, technological artifacts, as active mediators, affect the actions and experiences of humans, but they do not determine them. Some people have creatively circumvented the strict morality of earlier versions of the alcohol lock by having an air pump in the car (Vidal 2004). Nevertheless, these critiques underline the issues at stake in automating decision-making processes: computing can set constraints on the freedom a person has to act and thus affects the extent to which she can be held morally responsible.

The challenges that computer technologies present with regard to the conditions for ascribing responsibility indicate the limitations of conventional ethical frameworks in dealing with the question of moral responsibility. Traditional models of moral responsibility seem to be developed for the kinds of actions performed by an individual that have directly visible consequences (Waelbers 2009, Coeckelbergh 2009). However, in today's society attributions of responsibility to an individual or a group of individuals are intertwined with the artifacts with which they interact as well as with intentions and actions of other human agents that these artifacts mediate. Acting with computer technologies may require a different kind of analysis of who can be held responsible and what it means to be morally responsible. Below I discuss two ways in which scholars have taken up this challenge: 1) reconsidering what it means to be a moral agent and 2) reconsidering the concept of moral responsibility.

2. Can computers be moral agents?

Moral responsibility is generally attributed to moral agents and, at least in Western philosophical traditions, moral agency has been a concept exclusively reserved for human beings (Johnson 2001; Doorn and van de Poel 2012). Unlike animals or natural disasters, human beings in these traditions can be the originators of morally significant actions, as they can freely choose to act in one way rather than another way and deliberate about the consequences of this choice. And, although some people are inclined to anthropomorphize computers and treat them as if they were moral agents (Reeves and Nass 1996; Nass and Moon 2000; Rosenthal-von der Pütten 2013), most philosophers agree that current computer technologies should not be called moral agents, if that would mean that they could be held morally responsible. However, the limitations of traditional ethical vocabularies in thinking about the moral dimensions of computing have led some authors to rethink the concept of moral agency. It should be noted that some authors have also argued for a reconsideration of Western philosophical anthropocentric conceptions of moral patiency, in particular in regard to the question concerning the moral standing of artificial agents or robots (Floridi 2016, Gunkel 2020, Coeckelbergh 2020). The following will nevertheless focus on moral agency as these reflections on moral patiency tend to not address the challenges to moral responsibility.

2.1 Computers as morally responsible agents

The increasing complexity of computer technology and the advances in Artificial Intelligence (AI), challenge the idea that human beings are the only entities to which moral responsibility can or should be ascribed (Bechtel 1985; Kroes and Verbeek 2014). Dennett, for example, suggested that holding a computer morally responsible is possible if it concerned a higher-order intentional computer system (1997). An intentional system,

according to him, is one that can be predicted and explained by attributing beliefs and desires to it, as well as rationality. In other words, its behavior can be described by assuming the systems has mental states and that it acts according what it thinks it ought to do, given its beliefs and desires. At the time, Dennett noted that many computers were already intentional systems, but they lacked the higher-order ability to reflect on and reason about their mental states. They did not have beliefs about their beliefs or thoughts about desires. Dennett suggested that the fictional HAL 9000 that featured in the movie *2001: A Space Odyssey* would qualify as a higher-order intentional system that can be held morally responsible. Although advances in AI might not lead to HAL, he did see the development of computers systems with higher-order intentionality as a real possibility.

Sullins argues in line with Dennett that moral agency is not restricted to human beings (2006). He proposes that computers systems or, more specifically, robots are moral agents when they have a significant level of autonomy and they can be regarded at an appropriate level of abstraction as exhibiting intentional behavior. A robot, according to Sullins, would be significantly autonomous if it was not under the direct control of other agents in performing its tasks. Note that Sullins interprets autonomy in a narrow sense in comparison to the conception of autonomy in moral philosophy as property of human beings. He adds as a third condition that a robot also has to be in a position of responsibility to be a moral agent. That is, the robot performs some social role that carries with it some responsibilities and in performing this role the robot appears to have 'beliefs' about and an understanding of its duties towards other moral agents (p. 28). To illustrate what kind of capabilities are required for "full moral agency", he draws an analogy with a human nurse. He argues that if a robot was autonomous enough to carry out the same duties as a human nurse and had an understanding of its role and responsibilities in the health care systems, then it would be a "full moral agent". Sullins maintains that it will be some time before machines with these kinds of capabilities will

be available, but "even the modest robots of today can be seen to be moral agents of a sort under certain, but not all, levels of abstraction and are deserving of moral consideration" (p. 29).

Echoing objections to the early project of (strong) AI (Sack 1997),[3] critics of analyses such as presented by Dennett and Sullins, have objected to the idea that computer technologies can have capacities that make human beings moral agents, such as mental states, intentionality, common sense, emotion or empathy (Johnson 2006; Kuflik 1999; Nyholm 2018). They, for instance, point out that it makes no sense to treat computer system as moral agents that can be held responsible, for they cannot suffer and thus cannot be punished (Sparrow 2007; Asaro 2011). Veliz argues that computers may act like moral agents, but they lack sentience and are therefore 'moral zombies' (2021). Hakli and Makela argue that computers cannot have the kind of autonomy required for moral agency, because their capacities are a result of engineering and programming which undermines the autonomy of robots and disqualifies them as moral agents (2019). Or they argue, as Stahl does, that computers are not capable of moral reasoning, because they do not have the capacity to understand the meaning of the information that they process (2006). In order to comprehend the meaning of moral statements an agent has to be part of the form of life in which the statement is meaningful; it has to be able to take part in moral discourses. Similar to the debates about AI, critics continue to draw a distinction between humans and computers by noting various capacities that computers do not, and cannot, have that would justify the attribution of moral agency.

Some other critics do not contest that human beings might be able to build computer systems with the required capacities for moral agency, but question whether it is ethically appropriate to do so. Bryson, for instance, argues that even if it was possible to create artifacts with such capacities – and she assumes this might very well be possible – human beings have a

choice in the matter (2018). She defines a moral agent as "something deemed responsible by a society for its actions" (p. 16). Society can thus at one point deem it appropriate to view certain computer systems as moral agents, for instance as it would provide a short cut to figuring out how responsibility should be distributed. However, she argues that there is no necessary or predetermined position for these technologies in our society. This is because, she notes, computer technologies ethical frameworks are "artefacts of our societies, and therefore subject to human control" (p. 15). We can choose what capacities we equip these artefacts with, and she sees no coherent reason for creating artificial agents that human beings have to compete with in terms of moral agency or patiency.

2.2 Creating autonomous moral agents

In the absence of any definitive arguments for or against the possibility of future computer systems being morally responsible, researchers within the field of machine ethics aim to further develop the discussion by focusing instead on creating computer system that can behave *as if* they are moral agents (Moor 2006, Cervantes et al 2019 , Zoshak and Dew 2021). Research within this field has been concerned with the design and development of computer systems that can independently determine what the right thing to do would be in a given situation. According to Allen and Wallach, such *autonomous moral agents* (AMAs) would have to be capable of reasoning about the moral and social significance of their behavior and use their assessment of the effects their behavior has on sentient beings to make appropriate choices (2012; see also Wallach and Allen 2009 and Allen et al. 2000). Such abilities are needed, they argue, because computers are becoming more and more complex and capable of operating without direct human control in different contexts and environments. Progressively autonomous technologies already in development, such as military robots, driverless cars or trains and service robots in the home and for healthcare, will be involved in moral situations

that directly affect the safety and well-being of humans. An autonomous bomb disposal robot might in the future be faced with the decision which bomb it should defuse first, in order to minimize casualties. Similarly, a moral decision that a driverless car might have to make is whether to break for a crossing dog or avoid the risk of causing injury to the driver behind him. Such decisions require judgment. Currently operators make such moral decisions, or the decision is already inscribed in the design of the computer system. Machine ethics, Wallach and Allen argue, goes one step beyond making engineers aware of the values they build into the design of their products, as it seeks to build ethical decision-making into the machines.

To further specify what it means for computers to make ethical decisions or to put 'ethics in the machine', Moor distinguished between three different kinds of ethical agents: implicit ethical agents, explicit ethical agents, and full ethical agents (2006). The first kind of agent is a computer that has the ethics of its developers inscribed in their design. These agents are constructed to adhere to the norms and values of the contexts in which they are developed or will be used. Thus, ATM tellers are designed to have a high level of security to prevent unauthorized people from drawing money from accounts. An explicit ethical agent is a computer that can 'do ethics'. In other words, it can on the basis of an ethical model determine what would be the right thing to do, given certain inputs. The ethical model can be based on ethical traditions, such as Kantian, Confucianism, Ubuntu, or utilitarian ethics—depending on the preferences of its creators. These agents would 'make ethical decisions' on behalf of its human users (and developers). Such agents are akin to the autonomous moral agents described by Allen and Wallach. Finally, Moor defined full ethical agents as entities that can make ethical judgments and can justify them, much like human beings can. He claimed that although at the time of his writing there were no computer technologies that could be called fully ethical, it is an empirical question whether or not it would be possible in the future.

Few, if any, philosophers today would argue that this question has been answered in the postive.

The effort to build AMAs raises the question of how this effort affects the ascription of moral responsibility. As human beings would design these artificial agents to behave within pre-specified formalized ethical frameworks, it is likely that responsibility will still be ascribed to these human actors and those that deploy these technologies. However, as Allen and Wallach acknowledge, the danger of exclusively focusing on equipping robots with moral decision-making abilities, rather than also looking at the sociotechnical systems in which these robots are embedded, is that it may cause further confusion about the distribution of responsibility (2012). Robots with moral decision-making capabilities may present similar challenges to ascribing responsibility as other technologies, when they introduce new complexities that further obfuscate causal connections that lead back to their creators and users.

2.3 Expanding the concept of moral agency

The prospect of increasingly autonomous and intelligent computer technologies and the growing difficulty of finding responsible human agents lead Floridi and Sanders to take a different approach (2004). They propose to extend the class of moral agents to include artificial agents, while disconnecting moral agency and moral accountability from the notion of moral responsibility. They contend that "the insurmountable difficulties for the traditional and now rather outdated view that a human can be found accountable for certain kinds of software and even hardware" demands a different approach (p. 372). Instead, they suggest that artificial agents should be acknowledged as moral agents that can be held accountable, but not responsible. To illustrate they draw a comparison between artificial agents and dogs as sources of moral actions. Dogs can be the cause of a morally charged action, like damaging property or

helping to save a person's life, as in the case of search-and-rescue dogs. We can identify them as moral agents even though we generally do not hold them morally responsible, according to Floridi and Sanders: they are the source of a moral action and can be held morally accountable by correcting or punishing them.

Just like animals, Floridi and Sanders argue, artificial agents can be seen as sources of moral actions and thus can be held morally accountable when they can be conceived of as behaving like a moral agent from an appropriate *level of abstraction*. The notion of levels of abstraction refers to the stance one adopts towards and entity to predict and explain its behavior. At a low level of abstraction we would explain the behavior of a system in terms of its mechanical or biological processes. At a higher level of abstraction it can help to describe the behavior of a system in terms of beliefs, desires and thoughts. If at a high enough level a computational system can effectively be described as being interactive, autonomous and adaptive, then it can be held accountable according to Floridi and Sanders (p. 352). It, thus, does not require personhood or free will for an agent to be morally accountable; rather the agent has to act as if it had intentions and was able to make choices.

The advantage of disconnecting accountability from responsibility, according to Floridi and Sanders, is that it places the focus on moral agenthood, accountability and censure, instead of on figuring out which human agents are responsible. "We are less likely to assign responsibility at any cost, forced by the necessity to identify a human moral agent. We can liberate technological development of AAs [Artificial Agents] from being bound by the standard limiting view" (p. 376). When artificial agents 'behave badly' they can be dealt with directly, when their autonomous behavior and complexity makes it too difficult to distribute responsibility among human agents. Immoral agents can be modified or

deleted. It is then possible to attribute moral accountability even when moral responsibility cannot be determined.

Critics of Floridi's and Sanders' view on accountability and moral agency argue that placing the focus of analysis on computational artifacts by treating them as moral agents will draw attention away from the humans that deploy and develop them. Johnson, for instance, makes the case that computer technologies remain connected to the intentionality of their creators and users (2006). She argues that although computational artifacts are a part of the moral world and should be recognized as entities that have moral relevance, they are not moral agents, for they are not intentional. They are not intentional, because they do not have mental states or a purpose that comes from the freedom to act. She emphasizes that although these artifacts are not intentional, they do have intentionality, but their intentionality is related to their functionality. They are human-made artifacts and their design and use reflect the intentions of designers and users. Human users, in turn, use their intentionality to interact with and through the software. In interacting with the artifacts they activate the inscribed intentions of the designers and developers. It is through human activity that computer technology is designed, developed, tested, installed, initiated and provided with input and instructions to perform specified tasks. Without this human activity, computers would do nothing. Attributing independent moral agency to computers, Johnson claims, disconnects them from the human behavior that creates, deploys and uses them. It turns the attention away from the forces that shape technological development and limits the possibility for intervention. For instance, it leaves the issue of sorting out who is responsible for dealing with malfunctioning or immoral artificial agents or who should make amends for the harmful events they may cause. It postpones the question of who has to account for the conditions under which artificial agents are allowed to operate (Noorman 2009).

Yet, technologies can still be part of moral action, without being a moral agent. Several philosophers have stressed that moral responsibility cannot be properly understood without recognizing the active role of technology in shaping human action (Jonas 1984; Verbeek 2006; Johnson and Powers 2005; Nyholm 2018). Johnson, for instance, claims that although computers are not moral agents, the artifact designer, the artifact, and the artifact user should all be the focus of moral evaluation as they are all at work in an action (Johnson 2006). Humans create these artifacts and inscribe in them their particular values and intentions to achieve particular effects in the world and in turn these technological artifacts influence what human beings can and cannot do and affect how they perceive and interpret the world.

Similarly, Verbeek maintains that technological artifacts alone do not have moral agency, but moral agency is hardly ever 'purely' human. Moral agency generally involves a mediating artifact that shapes human behavior, often in way not anticipated by the designer (2008). Moral decisions and actions are co-shaped by technological artifacts. He suggests that in all forms of human action there are three forms of agency at work: 1) the agency of the human performing the action; 2) the agency of the designer who helped shaped the mediating role of the artifacts and 3) the artifact mediating human action. The agency of artifacts is inextricably linked to the agency of its designers and users, but it cannot be reduced to either of them. For him, then, a subject that acts or makes moral decisions is a composite of human and technological components. Moral agency is not merely located in a human being, but in a complex blend of humans and technologies.

In later papers, Floridi explores the concept of distributed moral actions (2013, 2016). He argues that some moral significant outcomes cannot be reduced to the moral significant actions of some individuals. Morally neutral actions of several individuals can still result in morally significant

events. Individuals might not have intended to cause harm, but nevertheless their combined actions may still result in moral harm to someone or something. In order to deal with the problem of subsequently assigning moral responsibility for such distributed moral actions, he argues that the focus of analysis should shift from the agents to the patients of moral actions. A moral action can then be evaluated in terms of the harm to the patient, regardless of the intentions of the agents involved. Assigning responsibility then focuses on whether or not an agent is causally accountable for the outcome and on adjusting their behavior to prevent harm. If the agents causally accountable - be they artificial or biological - are autonomous, can interact with each other and their environments and can learn from their interactions they can be held responsible for distributed moral actions, according to Floridi (2016).

3. Rethinking the concept of moral responsibility

In light of the noted difficulties in ascribing moral responsibility, several authors have critiqued the way in which the concept is used and interpreted in relation to computing. They claim that the traditional models or frameworks for dealing with moral responsibility fall short and propose different perspectives or interpretations to address some of the difficulties. Some of these will be discussed in this section.

3.1 Assigning responsibility

One approach is to rethink how moral responsibility is assigned (Gotterbarn 2001; Waelbers 2009). When it comes to computing practitioners, Gotterbarn identifies a potential to side-step or avoid responsibility by looking for someone else to blame. He attributes this potential to two pervasive misconceptions about responsibility. The first misconception is that computing is an ethically neutral practice. That is, according to Gotterbarn, the misplaced belief that technological artifacts

and the practices of building them are ethically neutral is often used to justify a narrow technology-centered focus on the development of computer system without taking the broader context in which these technologies operate into account. This narrow focus can have detrimental consequences. Gotterbarn gives the tragic case of a patient's death as a result of a faulty X-ray device as an example. A programmer was given the assignment to write a program that could lower or raise the X-ray device on a pole, after an X-ray technician set the required height. The programmer focused on solving the given puzzle, but failed to take account of the circumstances in which the device would be used and the contingencies that might occur. He, thus, did not consider the possibility that a patient could accidentally be in the way of the device moving up and down the pole. This oversight eventually resulted in a tragic accident. A patient was crushed by the device, when a technician set the device to tabletop height, not realizing that the patient was still underneath it. According to Gotterbarn, computer practitioners have a moral responsibility to consider such contingencies, even though they may not be legally required to do so. The design and use of technological artifacts is a moral activity and the choice for one particular design solution over another has real and material consequences.

The second misconception is that responsibility is only about determining blame when something goes wrong. Computer practitioners, according to Gotterbarn, have conventionally adopted a malpractice model of responsibility that focuses on determining the appropriate person to blame for harmful incidents (2001). This malpractice model leads to all sorts of excuses to shirk responsibility. In particular, the complexities that computer technologies introduce allow computer practitioners to side-step responsibility. The distance between developers and the effects of the use of the technologies they create can, for instance, be used to claim that there is no direct and immediate causal link that would tie developers to a malfunction. Developers can argue that their contribution to the chain of

events was negligible, as they are part of a team or larger organization and they had limited opportunity to do otherwise. The malpractice model, according to Gotterbarn, entices computer practitioners to distance themselves from accountability and blame.

The two misconceptions are based on a particular retrospective view of responsibility that places the focus on that which exempts one from blame and liability. In reference to Ladd, Gotterbarn calls this negative responsibility and distinguishes it from positive responsibility (see also Ladd 1989). Positive responsibility emphasizes "the virtue of having or being obliged to have regard for the consequences that his or her actions have on others" (Gotterbarn 2001, p. 227). Positive responsibility entails that part of the professionalism of computer experts is that they strive to minimize foreseeable undesirable events. It focuses on what ought to be done rather than on blaming or punishing others for irresponsible behavior. Gotterbarn argues that the computing professions should adopt a positive concept of responsibility, as it emphasizes the obligations and duties of computer practitioners to have regard for the consequences of one's actions and to minimize the possibility of causing harm. Computer practitioners have a moral responsibility to avoid harm and to deliver a properly working product, according to him, regardless of whether they will be held accountable if things turn out differently.

The emphasis on the positive moral responsibility of computer practitioners raises the question of how far this responsibility reaches, in particular in light of systems that many hands help create and the difficulties involved in anticipating contingencies that might cause a system to malfunction (Stieb 2008; Miller 2008). To what extent can developers and manufacturers be expected to exert themselves to anticipate or prevent the consequences of the use of their technologies or possible 'bugs' in their code? Computer systems today are generally incomprehensible to any single programmer and it seems unlikely that

complex computer systems can be completely error free. Martin argues in this respect that developers and the companies that decide to sell a computer technology *into* a particular context are responsible for the ethical implication of the use of these technologies in that context (2019). They are responsible for these implications because they are knowledgeable about the design decisions and are in a unique position to inscribe in the technology particular ideas (and biases) about what the technology should do and how it should do it. Thus, a company that creates and sells a risk-assessment system into the context of judicial decision-making is responsible for the ethical implications of biases resulting from its use and its opaqueness. The company willingly "takes on the obligation to understand the values of the decision to ensure the algorithms' ethical implications is congruent with the context" (p. 10). This, however, leaves open the question of what their responsibility is outside of that context. Should manufacturers of mobile phones have anticipated that their products would be used in roadside bombs? Manufacturers and their designers and engineers cannot foresee all the possible conditions under which their products will eventually operate. Moreover, how much control should a person have to be or feel responsible for the outcome of events? Such questions speak to what Santonio de Sio and Meccaci (2020) call the "active responsibility gap". They described active responsibility in much the same way as positive responsibility, in that it relates to the moral obligations of persons to ensure that the behavior of the systems they design, control, or use minimizes harm. A gap in this responsibility, according to them, results from these persons not being sufficiently aware, capable and motivated to see and act according to these obligations (p. 1059).

To address gaps in active responsibility (as well as backward-looking gaps in responsibility), Santonio de Sio and Meccaci suggest an approach that underlines the need to look at the broader sociotechnical system of human agents and technologies. Assigning responsibility requires looking at the

whole chain of design, development and use from a social, technical as well as organizational perspective. Each element in this change can be adjusted in an effort to address responsibility gaps, including the design of the computer systems as well as the organization that uses it. They base their approach on the idea of designing sociotechnial systems for *meaningful human control* as developed by Santonio de Sio and van den Hoven (2018). Meaningful human control is a concept that originally gained currency in the context of autonomous lethal weapons as an approach to addressing responsibility gaps. The question of what it means to have control over a technological system becomes particularly pertinent in situations where weapon systems are delegated tasks involved in target selection and engagement (Ekelhof 2019). Santoni di Sio and van den Hoven (2018) developed their conception of *meaningful human control* to get a more 'actionable analysis of control' that can help engineers, computing professionals, policy makers and designers think about how to design sociotechnical systems with responsibility in mind.

Santonio de Sio and van den Hoven assume that technology is part of the decisional mechanisms through which human agents carry out actions in the world and these mechanisms should be responsive to moral reasons for an agent to have control. Meaningful human control is thus conditional on the extent to which an outcome can be connected to the decisional mechanisms of human agents. To elucidate these connections, they formulate two necessary conditions for meaningful control called *tracking* and *tracing*.

Tracking requires that the whole sociotechnical system of technical, human and organizational elements should be responsive to moral reasons of the relevant agents and to the relevant facts of the circumstances. That is, the behaviour of the system should reflect the reasons, values, and intentions of these actors given particular circumstances. For example, if a machine learning system is trained to distinguish huskies from wolves, the

system should act according to the relevant reasons for doing this (e.g. alarming a farmer to the presence of a wolf). A machine learning system that is trained to make this distinction, but has only been shown pictures of wolves in the snow and huskies in more urban environments, it may deduce that the relevant distinguishing feature is snow. When shown a picture of a wolf in urban environment, it might subsequently misclassify the wolf as a husky. In this case, the system did not properly track the reasons of relevant human agents and the facts of the environment. Note that this tracking relation can involve multiple human agents along the chain. That is, the moral reasons do not necessarily have to come from the operator or user, they can also come from policy makers, designers or programmers.

Tracing requires that outcomes can be traced back to earlier decisions made by human agents that place them in the position resulting in the outcome. An example that the authors give is a drunk driver causing a serious accident. Even if the driver does not fulfill the conditions of responsibility at the time of the accident - because of mental incapacitation – the driver did make the earlier decision of drinking too much. The *tracing condition*, as Santoni di Sio and van den Hoven formulate it, assumes that it is possible that more than one human agent is involved in the actions that led to the outcome and that their actions are mediated by non-human systems. According to them, the tracing condition requires that the whole sociotechnical system is designed such that at least one human agent can have sufficient knowledge and moral awareness to be a potential target of legitimate response for the behavior of the system.

Santonio de Sio and van den Hoven's understanding of meaningful human control does not only have implications for the design of computer systems, but also for the design of the environment and the social and institutional practices. Tracking and tracing of relevant human moral reasons occurs on all these levels of design.

3.2 Responsibility as practice

Santonio de Sio and van den Hoven's analysis of meaningful human control draws attention to the social function of moral responsibility, which provides yet another perspective on the issue (Stahl 2006; Tigard 2021b). Both prospectively and retrospectively, responsibility works to organize social relations between people and between people and institutions. It sets expectations between people for the fulfillment of certain obligations and duties and provides the means to correct or encourage certain behavior. For instance, a robotics company is expected to build in safeguards that prevent robots from harming humans. If the company fails to live up to this expectation, it will be held accountable and in some cases it will have to pay for damages or undergo some other kind of punishment. The punishment or prospect of punishment can encourage the company to have more regard for system safety, reliability, sound design and the risks involved in their production of robots. It might trigger the company to take actions to prevent future accidents. Yet, it might also encourage it to find ways to shift the blame. The idea that responsibility is about interpersonal relationships and expectations about duties and obligations places the focus on the practices of holding someone responsible (Strawson 1962, Talbert 2022).

The particular practices and social structures that are in place to ascribe responsibility and hold people accountable, have an influence on how we relate to technologies. Just before the turn of the century, Nissenbaum already noted that the difficulties in attributing moral responsibility can, to a large extent, be traced back to the particular characteristics of the organizational and cultural context in which computer technologies are embedded. She argued that how we conceive of the nature, capacities and limitations of computing is of influence on the answerability of those who develop and use computer technologies (1997). At the time, she observed a systematic erosion of accountability in our increasingly computerized

society, where she conceived of accountability as a value and a practice that places an emphasis on preventing harm and risk. Accountability means there will be someone, or several people, to answer not only for the malfunctions in life-critical systems that cause or risk grave injuries and cause infrastructure and large monetary losses, but even for the malfunction that cause individual losses of time, convenience, and contentment (1994, p. 74). It can be used as "a powerful tool for motivating better practices, and consequently more reliable and trustworthy systems" (1997, p. 43). Holding people accountable for the harms or risks caused by computer systems provides a strong incentive to minimize them and can provide a starting point for assigning just punishment.

Cultural and organizational practices, at the time, however seemed to do the opposite, due to "the conditions under which computer technologies are commonly developed and deployed, coupled with popular conceptions about the nature, capacities and limitations of computing" (p. 43). Nissenbaum identified four barriers to accountability in society: (1) the problem of many hands, (2) the acceptance of computer bugs as an inherent element of large software systems, (3) using the computer as scapegoat and (4) ownership without liability. According to Nissenbaum people have a tendency to shirk responsibility and to shift the blame to others when accidents occur. The problem of many hands and the idea that software bugs are an inevitable by-product of complex computer systems are too easily accepted as excuses for not answering for harmful outcomes. People are also inclined to point the finger at the complexity of the computer and argue that "it was the computer's fault" when things go wrong. Finally, she perceived a tendency of companies to claim ownership of the software they developed, but to dismiss the responsibilities that come with ownership. To illustrate, she pointed to extended license agreements that assert a manufacturer's ownership of software, but disclaim any accountability for the quality or performance of the product.

These four barriers, Nissenbaum argued, stand in the way of a "culture of accountability" that is aimed at maintaining clear lines of accountability. Such a culture fosters a strong sense of responsibility as a virtue to be encouraged and everyone connected to an outcome of particular actions is answerable for it. Accountability, according to Nissenbaum, is different from liability. Liability is about looking for a person to blame and to compensate for damages suffered after the event. Once that person has been found, others can be let 'off the hook', which may encourage people to look for excuses, such as blaming the computer. Accountability, however, applies to all those involved. It requires a particular kind of organizational context, one in which answerability works to entice people to pay greater attention to system safety, reliability and sound design, in order to establish a culture of accountability (see also Martin 2019, Herkert et AL 2020) . An organization that places less value on accountability and that has little regards for responsibilities in organizing their production processes is more likely to allow their technological products to become incomprehensible. Nissenbaum's analysis illustrates that our practices of holding someone responsible - the established ways of holding people to account and of conveying expectations about duties and obligations - are continuously changing and negotiated, partly as a response to the introduction of new technologies (see also Noorman 2012).

A lack of such a culture of accountability can lead to responsibility being attributed to the wrong people. These people become, what Madeleine Clare Elish calls the moral crumple zone (2019). These human actors absorb responsibility, even though they only have very limited or no control over the systems they work with. Elish argues that, given the complexity of technological systems, the media and the public tend to blame accidents on human error and misattribute responsibility to the nearest human operator, such as a pilot or maintenance personnel, rather than the technological systems or the decision-makers higher up the chain.

Nissenbaum argued that the context in which technologies are developed and used has a significant influence on the ascription of moral responsibility, but several authors have stressed that moral responsibility cannot be properly understood without recognizing the active role of technology in shaping human action (Jonas 1984; Verbeek 2006; Johnson and Powers 2005; Waelbers 2009). According to Johnson and Powers it is not enough to just look at what humans intend and do. "Ascribing more responsibility to persons who act with technology requires coming to grips with the behavior of the technology" (2005, p. 107). One has to consider the various ways in which technological artifacts mediate human actions. Moral responsibility is, thus, not only about how the actions of a person or a group of people affect others in a morally significant way; it is also about how their actions are shaped by technology. Moral responsibility from this perspective is not located in an individual or an interpersonal relationship, but is distributed among humans and technologies.

4. Conclusion

Computer technologies have challenged conventional conceptions of moral responsibility and have raised questions about how to distribute responsibility appropriately. Can human beings still be held responsible for the behavior of complex computer technologies that they have limited control over or understanding of? Are human beings the only agents that can be held morally responsible or can the concept of moral agent be extended to include artificial computational entities? In response to such questions philosophers have reexamined the concepts of moral agency and moral responsibility. Although there is no clear consensus on what these concepts should entail in an increasingly digital society, what is clear from the discussions is that any reflection on these concepts will need to address how these technologies affect human action and where responsibility for action begins and ends.

Bibliography

Allen, C. & W. Wallach, 2012. "Moral Machines. Contradiction in Terms or Abdication of Human Responsibility?" in P. Lin, K. Abney, and G. Bekey (eds.), *Robot ethics. The ethics and social implications of robotics*. Cambridge, Massachusetts: MIT Press.

Allen, C., G. Varner & J. Zinser, 2000. "Prolegomena to any Future Artificial Moral Agent," *Journal of Experimental and Theoretical Artificial Intelligence*, 12: 251–261.

Allen, C. W. Wallach & I. Smit, 2006. "Why Machine Ethics?" *Intelligent Systems, IEEE* , 21(4):12–17.

Angwin, J., J. Larson, S. Mattu & L. Kirchner, 2016. "Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks", *ProPublica*, May 23, 2016,Angwin et al. 2016 available online

Asaro, P., 2011. "A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics," in P. Lin, K. Abney, and G. Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*, Cambridge, MA: MIT Press.

Bechtel, W., 1985. "Attributing Responsibility to Computer Systems," *Metaphilosophy*, 16(4): 296–306

Bijker, W. E., T. P. Hughes, & T. Pinch, 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, London, UK: The MIT Press.

Bovens, M. & S. Zouridis, 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.

Boyd, D. & K. Crawford, 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly

Phenomenon," *Information, Communication, & Society* 15(5): 662–679.

Bryson, J.J., 2018. "Patiency is not a virtue: the design of intelligent systems and systems of ethics," *Ethics and Information Technology*, 20(1): 15-26.

Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F. and Ramos, F., 2020."Artificial moral agents: A survey of the current status," *Science and Engineering Ethics*, 26(2): 501-532.

Coeckelbergh, M., 2009. "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents," *AI & Society*, 24: 181–189.

——, 2012. "Moral responsibility, technology, and experiences of the tragic: From Kierkegaard to offshore engineering," *Science and Engineering Ethics*, 18(1): 35-48.

——, 2013. "Drones, information technology, and distance: mapping the moral epistemology of remote fighting," *Ethics and Information Technology*, 15(2): 87-98.

Coeckelbergh, M., 2020. "Artificial intelligence, responsibility attribution, and a relational justification of explainability," *Science and Engineering Ethics*, 26(4): 2051-2068.

Coeckelbergh, M. & R. Wackers, 2007. "Imagination, Distributed Responsibility and Vulnerable Technological Systems: the Case of Snorre A," *Science and Engineering Ethics*, 13(2): 235–248.

Cummings, M. L., 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems," published online: 19 Jun 2012, American Institute of Aeronautics and Astronautics. doi:10.2514/6.2004-6313

Diakopoulos, N., 2020. "Transparency". In M. Dubber, F. Pasquale, & S. Das (Eds.), *Oxford handbook of ethics and AI* (pp. 197–214). Oxford University Press.

Dennett, D. C., 1997. "When HAL Kills, Who's to Blame? Computer Ethics," in *HAL's Legacy: 2001's Computer as Dream and Reality*, D. G. Stork (ed.), Cambridge, MA: MIT Press.

Denning, P. J., 1989. "The Science of Computing: The Internet Worm," *American Scientist*, 77(2): 126–128.

Dignum, V., 2020. "Responsibility and artificial intelligence," *The Oxford Handbook of Ethics of AI*, Markus D. Drubber et al. (eds.), Oxford: Oxford University Press, pp. 214-231.

Doorn, N. & van de Poel, I., 2012. "Editors Overview: Moral Responsibility in Technology and Engineering," *Science and Engineering Ethics*, 18: 1–11.

Ekelhof, M., 2019. "Moving beyond semantics on autonomous weapons: Meaningful human control in operation," *Global Policy*, 10(3): 343-348.

Elish, M.C., 2019. "Moral crumple zones: Cautionary tales in human-robot interaction," *Engaging Science, Technology, and Society*, 5: 40-60.

Eshleman, A., 2016. "Moral Responsibility," in *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/>

Eubanks, V., 2018. "Automating inequality: How high-tech tools profile, police, and punish the poor," New York: St. Martin's Press.

Felt, U., Fouché, R., Miller, C. A., & Smith-Doerr, L., 2016. *The Handbook of Science and Technology Studies*, Cambridge, MA: MIT Press.

Fisher, J. M., 1999. "Recent work on moral responsibility," *Ethics*, 110(1): 93–139.>

Floridi, L., & J. Sanders, 2004. "On the Morality of Artificial Agents," *Minds and Machines*, 14(3): 349–379.

Floridi, L., 2013. "Distributed morality in an information society," *Science and Engineering Ethics*, 19(3): 727–743.

——, 2016. "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions," *Philosophical Transactions of the Royal Society A* (Mathematical Physical and Engineering Sciences), 374(2083); doi: 10.1098/rsta.2016.0112

Friedman, B., 1990. "Moral Responsibility and Computer Technology," Institute of Education Sciences ERIC Number ED321737, [Friedman 1990 available online].

—— (ed.), 1997. *Human Values and the Design of Computer Technology*, Stanford: CSLI Publications; New York: Cambridge University Press.

Gorwa, R., Binns, R., & Katzenbach, C., 2020. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, 7(1); first online 28 February 2020. doi:10.1177/2053951719897945

Gotterbarn D., 2001. "Informatics and professional responsibility," *Science and Engineering Ethics*, 7(2): 221–230.

Graubard, S. R., 1988. *The Artificial Intelligence Debate: False Starts, Real Foundations*, Cambridge, MA: MIT Press.

Gray, C. H., 1997. "AI at War: The Aegis System in Combat," *Directions and Implications of Advanced Computing*, D. Schuler, (ed.), New York: Ablex, pp. 62–79.

Gunkel, D. J., 2020. "A vindication of the rights of machines," in *Machine Ethics and Robot Ethics*, W. Wallach and P. Asaro (eds.), London: Routledge, pp. 511-530.

Hakli, R. and Mäkelä, P., 2019. "Moral responsibility of robots and hybrid agents," *The Monist*, 102(2): 259-275.

Hart, H. L. A., 1968. *Punishment and Responsibility*, Oxford: Oxford University Press.

Herkert, J., Borenstein, J., & Miller, K., 2020. "The Boeing 737 MAX: Lessons for engineering ethics". *Science and engineering ethics*, *26*, 2957-2974.

Hughes, T.P., 1987. "The evolution of Large Technological System," in W. E. Bijker, T. P. Hughes, & T. Pinch (eds.), *The Social Construction of Technological Systems*, Cambridge, MA: The MIT Press, pp. 51–82.

IJsselsteijn, W., Y. de Korte, C. Midden, B. Eggen, & E. Hoven (eds.), 2006. *Persuasive Technology*, Berlin: Springer-Verlag.

Johnson, D. G., 2001. *Computer Ethics*, 3rd edition, Upper Saddle River, New Jersey: Prentice Hall.

——, 2006. "Computer Systems: Moral Entities but not Moral Agents," *Ethics and Information Technology*, 8: 195–204.

Johnson, D. G. & T. M. Power, 2005. "Computer systems and responsibility: A normative look at technological complexity," *Ethics and Information Technology*, 7: 99–107.

Jonas, H., 1984. *The Imperative of Responsibility. In search of an Ethics for the Technological Age*, Chicago: The Chicago University Press.

Kroes, P.& P.P. Verbeek (eds.), 2014. *The Moral Status of Technical Artefacts*, Dordrecht: Springer

Kuflik, A., 1999. "Computers in Control: Rational Transfer of Authority or Irresponsible Abdication of Authority?" *Ethics and Information Technology*, 1: 173–184.

Ladd. J., 1989. "Computers and Moral Responsibility. A Framework for an Ethical Analysis," in C.C. Gould (ed.), *The Information Web. Ethical and Social Implications of Computer Networking*, Boulder, Colorado: Westview Press, pp. 207–228.

Latour, B., 1992. "Where are the Missing Masses? The Sociology of a Few Mundane Artefacts," in W. Bijker & J. Law (eds.), *Shaping Technology/Building Society: Studies in Socio-Technical Change*, Cambridge, Massachusetts: The MIT press, pp. 225–258.

Leveson, N. G. & C. S. Turner, 1993. "An Investigation of the Therac-25 Accidents," *Computer*, 26(7): 18–41.

Leveson, N., 1995. "Medical Devices: The Therac-25," in N. Leveson, *Safeware. System, Safety and Computers*, Boston: Addison-Wesley.

Martin, K., 2019. "Ethical implications and accountability of algorithms," *Journal of Business Ethics*, 160(4): 835-850.

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

McCorduck, P., 1979. *Machines Who Think*, San Francisco: W.H. Freeman and Company.

Miller, K. W., 2008. "Critiquing a critique," *Science and Engineering Ethics*, 14(2): 245–249.

Moor, J.H., 2006. "The Nature, Importance, and Difficulty of Machine Ethics," *Intelligent Systems* (IEEE), 21(4): 18–21.

Nissenbaum, H., 1994. "Computing and Accountability," *Communications of the Association for Computing Machinery*, 37(1): 72–80.

——, 1997. "Accountability in a Computerized Society," in B. Friedman (ed.), *Human Values and the Design of Computer Technology*, Cambridge: Cambridge University Press, pp. 41–64.

Nass, C. & Y. Moon, 2000. "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, 56(1): 81–103.

Noorman, M., 2009. *Mind the Gap: A Critique of Human/Technology Analogies in Artificial Agents Discourse*, Maastricht: Universitaire Pers Maastricht.

——, 2012. "Responsibility Practices and Unmanned Military Technologies," *Science and Engineering Ethics*, 20(3): 809–826.

Nihlén Fahlquist, J., Doorn, N., &Van de Poel, I., 2015. "Design for the value of responsibility," in *Handbook of ethics, values and technological design : Sources, Theory, Values and Application*

Domains, Jeroen van den Hoven, Ibo van de Poel and Pieter Vermaas (eds.), Dordrecht: Springer.

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

Parasuraman, R. & V. Riley, 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors: the Journal of the Human Factors Society*, 39(2): 230–253.

Pasquale, F., 2015. *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Polder-Verkiel, S. E., 2012. "Online responsibility: Bad samaritanism and the influence of internet mediation," *Science and engineering ethics*, 18(1): 117-141.

Ravenscraft, E., (2020). "How to Spot—and Avoid—Dark Patterns on the Web", *Wired*, July 29, Ravenscraft 2020 available online.

Reeves, B. & C. Nass, 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge: Cambridge University Press.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C., 2013. "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, 5(1): 17–34.

Sack, W., 1997. "Artificial Human Nature," *Design Issues*, 13: 55–64.

Santoni de Sio, F. and Van den Hoven, J., 2018. "Meaningful human control over autonomous systems: A philosophical account," *Frontiers in Robotics and AI*, 5, first online 28 February 2018. doi:10.3389/frobt.2018.00015

Santoni de Sio, F., & Mecacci, G., 2021. "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy & Technology*, 34: 1057–1084.

Sartor, G. and M. Viola de Azevedo Cunha, 2010. "The Italian Google-Case: Privacy, Freedom of Speech and Responsibility of Providers for User-Generated Contents," *International Journal of Law and Information Technology*, 18(4): 356–378.

Searle, J. R., 1980. "Minds, brains, and programs" *Behavioral and Brain Sciences*, 3(3): 417–457.

Singel, R., 2010. "Does Italy's Google Conviction Portend More Censorship?" *Wired* (February 24th, 2010), Singel 2010 available online.

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

Stahl, B. C., 2004. "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents," *Minds and Machines*, 14: 67–83.

——, 2006. "Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency," *Ethics and Information Technology*, 8: 205–213.

Stieb, J. A., 2008. "A Critique of Positive Responsibility in Computing," *Science and Engineering Ethics*, 14(2): 219–233.

Strawson, P., 1962. "Freedom and Resentment," in *Proceedings of the British Academy*, 48: 1-25.

Suchman, L., 1998. "Human/machine reconsidered," *Cognitive Studies*, 5(1): 5–13.

Sullins, J. P., 2006. "When is a Robot a Moral Agent?" *International Review of Information Ethics*, 6(12): 23–29.

Swierstra, T., Waelbers, K., 2012. "Designing a Good Life: A Matrix for the Technological Mediation of Morality," *Science and Engineering Ethics*, 18: 157–172. doi:10.1007/s11948-010-9251-1

Taddeo, M. and L. Floridi, 2015. "The Debate on the Moral Responsibilities of Online Service Providers," *Science and Engineering Ethics*, 22(6): 1575–1603.

Talbert, Matthew, 2022. "Moral Responsibility," *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.) URL = < href="https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/">https://plato.stanford.edu/archives/win2016/entries/m responsibility/>.

Tigard, D. W., 2021a. "Responsible AI and moral responsibility: a common appreciation," *AI and Ethics*, 1(2): 113-117.

——, 2021b. "Artificial moral responsibility: How we can and cannot hold machines responsible," *Cambridge Quarterly of Healthcare Ethics*, 30(3): 435-447.

U.S. Department of Defense, 2009. "FY2009–2034 Unmanned Systems Integrated Roadmap," available online .

Van den Hoven, J., 2002. "Wadlopen bij Opkomend Tij: Denken over Ethiek en Informatiemaatschappij," in J. de Mul (ed.), *Filosofie in Cyberspace*, Kampen: Uitgeverij Klement, pp. 47–65.

Véliz, C., 2021. "Moral zombies: why algorithms are not moral agents," *AI & Society*, 36: 487–497. doi:10.1007/s00146-021-01189-x

Verbeek, P. P., 2006. "Materializing Morality: Design Ethics and Technological Mediation," *Science, Technology and Human Values*, 31(3): 361–380.

Verbeek, P. P., 2021. *What Things Do*, University Park, PA: Pennsylvania State University Press.

Vidal, J., 2004. "The alco-lock is claimed to foil drink-drivers. Then the man from the Guardian had a go …," *The Guardian*, August 5th, 2004.

Waelbers, K., 2009. "Technological Delegation: Responsibility for the Unintended," *Science & Engineering Ethics*, 15(1): 51–68.

Wallach, W. and C. Allen, 2009. *Moral Machines. Teaching Robots Right from Wrong*, Oxford: Oxford University Press.

Whitby, B., 2008. "Sometimes it's hard to be a robot. A call for action on the ethics of abusing artificial agents," *Interacting with Computers*, 20(3): 326–333.

John Zerilli; John Danaher; James Maclaurin; Colin Gavaghan; Alistair Knott; Joy Liddicoat; Merel Noorman, 2021. "7 Autonomy," in *A Citizens Guide to Artificial Intelligence*, Cambridge, MA: MIT Press, pp. 107-126.

Zuboff, S., 1982. "Automate/Informate: The Two Faces of Intelligent Technology," *Organizational Dynamics*, 14(2):5–18

Academic Tools

- How to cite this entry.
- Preview the PDF version of this entry at the Friends of the SEP Society.
- Look up topics and thinkers related to this entry at the Internet Philosophy Ontology Project (InPhO).
- Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

Journals On-line

- Ethics and Information Technology: A peer-reviewed journal dedicated to advancing the dialogue between moral philosophy and the field of information and communication technology (ICT).
- Science and Engineering Ethics: Science and Engineering Ethics is a multi-disciplinary journal that explores ethical issues of direct concern to scientists and engineers.
- Philosophy and Technology: A journal that addresses the expanding scope and unprecedented impact of technologies, in order to improve

the critical understanding of the conceptual nature and practical consequences, and hence provide the conceptual foundations for their fruitful and sustainable developments.
- Big Data and Society: A peer-reviewed scholarly journal that publishes interdisciplinary work principally in the social sciences, humanities and computing and their intersections with the arts and natural sciences about the implications of Big Data for societies.

Centers

- 4TU Centre for Ethics and Technology .
- Computer Professionals for Social Responsibility (CPSR)

Organizations

- IACAP: International Association for Computing and Philosophy: concerned with computing and philosophy broadly construed, including the use of computers to teach philosophy, the use of computers to model philosophical theory, as well as philosophical concerns raised by computing.
- Responsible Robotics : Organization to promote the responsible design, development, implementation, and policy of robots embedded in society.
- Moral Machine : A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

Blogs

- Moral machines: blog on the theory and development of artificial moral agents and computational ethics.

Related Entries

information technology: and moral values | information technology: phenomenological approaches to ethics and | moral responsibility | technology, philosophy of

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. SES 1058457.

Notes to Computing and Moral Responsibility

1. The term technological artifacts here refers to the (socially) constructed material or physical objects such as computers, cars and refrigerators, that human beings create and use to achieve a particular purpose or goal. This conception of technological artifacts is often used in social and historical studies to distinguish artifacts from natural objects and other socially constructed artifacts like regulatory laws (Hughes 1982; Bijker et al. 1987). For more on the concept of artifacts see the entry Artifacts.

2. According to Bijker et al. interpretive flexibility of technological artifacts means that "there is flexibility in how people think of, or interpret, artefacts" and "that there is flexibility in how artefacts are designed" (Bijker et al. 1995, p. 40). That is, different 'relevant social groups' have varying criteria for judging what makes a design superior or even workable, depending on, often competing, goals and interests, as well as on distinct ideas about what a particular artifact should do.

3. A long running philosophical debate about Artificial Intelligence is centered on the thesis that processes of the mind could be generated by computational structures (McCorduck 1979). Critics of AI have taken

exception to the suggestion that the human mind and computers could be thought of as governed by the same general principles (Graubard 1988). They have argued against the presupposition that knowledge and intelligence could be captured in computational structures and mathematical or logical models. These critics have provided a range of proposed inherent properties or abilities that humans have and machines lack, such as emotion, common sense and intentionality. One of these critics, Searle, was the first to use the term 'strong AI' to refer to the philosophical position that a computer with the right kind of programs can literally be a mind that is able to understand and have other cognitive states (Searle 1980). He distinguished this kind of research from, what he called, 'weak AI'. Weak AI makes no claims about computers being minds and merely argues that computers are useful for testing particular explanations of processes of the mind because they simulate these processes. Contrary to strong AI, this position does not claim, according to Searle, that computers literally *are* the explanation (see also the entry on the Chinese Room argument).