Computing and Moral Responsibility

First published Wed Jul 18, 2012

Traditionally philosophical discussions on moral responsibility have focused on the human components in 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; Waelbers 2009).[1] As we interact with and through these artifacts, they affect the decisions that we make and how we make them (Latour 1992). 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 change the character of human action and as a result it challenges conventional notions about how to distribute moral responsibility (Jonas 1984; Johnson 2001).

Computing presents a particular case for understanding the role of technology in moral responsibility. As these technologies 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 further complicate 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? Who is accountable when electronic records are lost or when they contain errors? 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 consequences. 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 morally responsibility establishes a link between a person or a group of people (the subject) and someone or something (the object) that is affected by the actions of the subject. This 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, it is not always clear when the ascription of moral responsibility is appropriate. 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). 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 subject and an object 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 contingent ways. 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).

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 (Eshelman 2009; 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 she 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 incident 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 contribution of multiple actors in the development and deployment of technologies is known as the problem of 'many hands' (Friedman 1990; Nissenbaum 1994; Jonas 1984). One much-discussed 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 1980's 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. This is not to say that there is no moral responsibility in this case (Nissenbaum 1994; Gotterbarn 2001), as many actors could have acted differently, but it makes it 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 remote-controlled drone over Afghanistan from their ground-control station in the United States. These technologies enable people to act over greater distances, but this remoteness can dissociate the original actions from its eventual consequences (Waelbers 2009). 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 Internet filter, often at the request of a manager or a client, with the aim of shielding particular content from its 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). 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, technological artifacts do not determine human action. 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).[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 her actions. She has be aware of the possible risks or harms that her actions might cause. It is unfair to hold someone responsible for something if they could not have 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).

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). Users only see part of the many computations that a computer performs and are for the most part are 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 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). 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. She tended 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 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. “Computer 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). The social or legal conventions that govern these new modes of conduct take some time to emerge and the initial absence of these conventions contributes to confusion about responsibilities. For example, the ability for users to upload and share text, videos and images publicly on the Internet raises a whole new 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 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 and viruses. 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. 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. 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).

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 (Eshleman 2009)?

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 few decades progressively standardized and formalized their production processes. The process of issuing decisions about student loans, speeding tickets or tax returns is carried out almost entirely 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. 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). 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.

The automation of bureaucratic processes illustrates that some computer technologies are intentionally designed to limit the discretion of some human beings. Indeed the relatively new field of Persuasive Technology explicitly aims to develop technological artifacts that persuade humans to perform in 'desirable' ways (IJsselsteijn et al. 2006). 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 seducing users (Verbeek 2006). 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.

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 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). 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.

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). 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 moral agents (Reeves and Nass 1996), most philosophers agree that current computer technologies should not be called moral agents, if that would mean that they could be held morally responsibility. 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.

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). Dennett, for example, suggests 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. Many computers today, according to Dennett, are already intentional systems, but they lack the higher-order ability to reflect on and reason about their mental states. They do not have beliefs about their beliefs or thoughts about desires. Dennett suggests 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 current advances in AI might not lead to HAL, he does see the development of computers systems with higher-order intentionality as a real possibility.

Sullins argues in line with Dennett that moral agency does not require personhood (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. However, 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 on offer, 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 or emotion (Johnson 2006; Kuflik 1999). 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). 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.

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). 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 distinguishes 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 traditional ethical theories, such as Kantian 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 defines full ethical agents as entities that can make ethical judgments and can justify them, much like human beings can. He claims that although there are no computer technologies today that can be called fully ethical, it is an empirical question whether or not it would be possible in the future.

The effort to build AMAs raises the question of how this effort affects the ascription of moral responsibility. If these technologies are not moral agents like human beings are, can they be held morally responsible? 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 the chains of accountability 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 do not have mental states or a purpose that comes from the freedom to act. She emphasizes that these artifacts 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, to say that technologies are not moral agents is not to say that they are not part of moral action. 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; Waelbers 2009). 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, building on the work of Bruno Latour, he argues that 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.

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.

One approach is to rethink how moral responsibility is assigned (Gotterbarn 2001; Waelbers 2009). When it comes to computing practitioners, Gotterbarn observes a tendency to side-step or avoid responsibility by looking for someone else to blame. He attributes this tendency to two pervasive misconceptions about responsibility. The first misconception is that computing is an ethically neutral practice. According to Gotterbarn this 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 example of a programmer who was given the assignment to write a program that could lower or raise an 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. 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 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 prospective 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? These systems are generally incomprehensible to any single programmer and it seems unlikely that complex computer systems can be completely error free. Moreover, designers and engineers cannot foresee all the possible conditions under which their products will eventually operate. Should manufacturers of mobile phones have anticipated that their products would be used in roadside bombs? A more fundamental question is whether computer programmers have a broader responsibility to the welfare of the public or that they are primarily responsible for performing their tasks well?

Nevertheless, the distinction between positive and negative responsibility underlines that holding someone morally responsible has a function, which provides yet another perspective on the issue (Stahl 2006). 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 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. Nissenbaum contends 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 computers technologies are embedded. She argues 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). She observes a systematic erosion of accountability in our increasingly computerized society, where she conceives 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.

Current cultural and organizational practices however 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 identifies four barriers to accountability in today's 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 perceives a tendency of companies to claim ownership of the software they develop, but to dismiss the responsibilities that come with ownership. Current day computer programs come with extended license

agreements that assert the manufacturer's ownership of the software, but disclaim any accountability for the quality or performance of the product. They also dismiss any liability for the consequential damages resulting from defects in the software.

These four barriers, Nissenbaum holds, 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. 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 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" (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.

Bibliography

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

- 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.
- 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.” Paper presented at the AIAA 1st Intelligent Systems Technical Conference, Chicago.
- 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.
- Eshleman, A.. 2009. “Moral Responsibility,“ in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 Edition), URL = <http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/>.
- 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.
- Friedman, B. 1990. “Moral Responsibility and Computer Technology.” Paper Presented at the Annual Meeting of the American Educational Research Association, Boston, Massachusetts.
- —— (ed.). 1997. *Human Values and the Design of Computer Technology,* Stanford: CSLI Publications; NY: Cambridge University Press
- 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 Massachusetts: The MIT Press.
- Gray, C. H.. 1997. “AI at War: The Aegis System in Combat,” *Directions and Implications of Advanced Computing 1990,* Vol. III, D. Schuler, (ed.), NY: Ablex, pp. 62–79.
- Hart, H. L. A.. 1968. *Punishment and Responsibility*. Oxford: Oxford University Press.
- 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*, 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* (3 ed.). 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.
- 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. and 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*, Addison-Wesley.
- McCorduck, P.. 1979. *Machines who Think*. San Francisco, US: 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.
- Noorman, M.. 2009. *Mind the gap a critique of human/technology analogies in artificial agent discourse*. Maastricht, the Netherlands: Universitaire Pers Maastricht.
- 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.
- 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.
- Sack, W.. 1997. “Artificial Human Nature,” *Design Issues*, 13: 55–64.
- 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) [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.
- 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.

- US Department of Defense [US DoD]. 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, the Netherlands: Uitgeverij Klement, pp. 47–65
- Verbeek, P. P.. 2006. “Materializing Morality,” *Science, Technology and Human Values*, 31(3): 361–380.
- 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, UK: 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.
- Zuboff, S.. 1982. “Automate/Informate: The Two Faces of Intelligent Technology,”*Organizational Dynamics*, 14(2): 5–18