12 Ethics of Robotics

Vincent C. Müller

12.1 Introduction

This chapter will provide a comprehensive introduction to the ethics of robotics, with a particular emphasis on the integration of artificial intelligence (AI) and robotics. After the introduction to the field in section 12.1, the main themes are, in section 12.2, ethical issues that arise with robotics systems as *objects* (i.e., tools made and used by humans), where the main sections are privacy, human-robot interaction, employment, and the effects of autonomy, and in section 12.3, robotics systems as *subjects* (i.e., when ethics is for the systems themselves in machine ethics and artificial moral agency). Many of these questions concern the use of AI, so the ethics of AI will play a role in this chapter.

For each section within these themes, we provide a general explanation of the ethical issues, we outline existing positions and arguments, and then we analyze how this plays out with current technologies and finally what policy consequences may be drawn.

12.1.1 Background of the Field

The ethics of robotics is often focused on "concerns" of various sorts—which is a typical response to new technologies. The task of an essay such as this is to analyze the issues and to deflate the nonissues. Some technologies, such as nuclear power, cars, or plastics, have caused ethical and political discussion and significant policy efforts to control the trajectory of these technologies—usually once some damage is done.

The ethics of robotics has seen significant press coverage in recent years, which supports this kind of work but also may end up undermining it: It often talks as though we already knew what would be ethical and as if the issues were just what future technology will bring and what we should do about it. Press coverage thus focuses on considerations of risk, security (Brundage et al. 2018), and the prediction of impact (e.g., on the job market). The result is a discussion of essentially technical problems and on how to achieve the desired outcome. Another result is that much of the current discussion in policy and industry, with its focus on image and public relations—where the label "ethical" is really not much more than the new "green," is perhaps used for "ethics washing." For a problem to qualify as a problem for robot ethics would require that we do *not* readily know what is the right thing to do. In this sense, job loss, theft, or killing with a robot are not a problem for ethics, but whether these are permissible under certain circumstances *is* such a problem.

A last caveat is in order for our presentation: The ethics of robotics is a very young field within applied ethics, with significant dynamics but few well-established issues and no authoritative overviews—though surveys for the ethics of robotics exist (Lin, Abney, and Jenkins 2017; Royakkers and van Est 2016; Calo, Froomkin, and Kerr 2016; Tzafestas 2016; European Group on Ethics in Science and New Technologies 2018). So this article cannot just reproduce what the community has achieved thus far but must propose an ordering where little order exists.

12.1.2 A Note on Policy

There is significant public discussion about robot ethics, and there are frequent pronouncements from politicians that the matter requires new policy, but actual technology policy is difficult to plan and to enforce. It can take many forms, from incentives and funding, infrastructure, taxation, or good-will statements to regulation by various actors and the law. Policy for robotics will possibly come into conflict with other aims of technology policy or general policy. One important practical aspect is which agents are involved in the development of a policy and what power structures oversee it.

For people who work in ethics and policy, there is probably a tendency to overestimate the impact and the threats from a new technology and to underestimate how far current regulation can reach (e.g., for product liability). On the other hand, for businesses, the military, and some administrations there is an interest to "talk" and to preserve a good public image but not to "do" anything. Governments, parliaments, associations, and industry circles in industrialized countries have produced reports and white papers in recent years, and some have generated good-will slogans. For a survey, see (Jobin, Ienca, and Vayena 2019).

Though very little actual policy has been produced, there are some notable beginnings. The latest EU policy document suggests "trustworthy AI" should be lawful, ethical, and technically robust and then spells this out as seven requirements: human oversight, technical robustness, privacy and data governance, transparency, fairness, well-being, and accountability (AI HLEG 2019). Much European research now runs under the slogan of "responsible research and innovation" (RRI), and "technology assessment" has been a standard field since the advent of nuclear power. Professional ethics is also a standard field in information technology, and this includes issues that are relevant here. We also expect that much policy will eventually cover specific uses or technologies of robotics, rather than the field as a whole (see Calo 2018; Stahl, Timmermans, and Mittelstadt 2016; Johnson and Verdicchio 2017; Giubilini and Savulescu 2018; Crawford and Calo 2016). The more political angle of technology is often discussed in "science and technology studies" (STS). As books like *The Ethics of Invention* (Jasanoff 2016) show, the concerns are often quite similar to those of ethics (Jacobs et al. 2019).

12.2 Ethics for the Use of Robotics Systems

In this section we outline the ethical issues of the human use of AI and robotics systems that can be more or less autonomous—which means we look at issues that arise with certain uses but would not arise with others. It must be kept in mind, however, that the design of technical

artifacts has ethical relevance for their use (Houkes and Vermaas 2010; Verbeek 2011), so beyond "responsible use," we also need "responsible design" in this field.

12.2.1 Human-Robot Interaction

Human-robot interaction (HRI) now pays significant attention to ethical matters, to the dynamics of perception from both sides, and to the different interests and the intricacy of the social context, including coworking (e.g., Arnold and Scheutz 2017).

Deception and authenticity

The central questions here often involve whether a robot involves deception, or perhaps violates human dignity or the Kantian requirement of "respect for humanity" (Lin, Abney, and Jenkins 2017). Humans very easily attribute mental properties to objects, and empathize with them, especially when the outer appearance of these objects is similar to that of living beings. This can be used to deceive humans (or animals) into attributing more intellectual or even emotional significance to robots than they deserve. Some parts of humanoid robotics are problematic in this regard (e.g., Hiroshi Ishiguro's remote-controlled Geminoids), and there are cases that have clearly been deceptive for public relations purposes (e.g., Hanson Robotics' "Sophia," with exaggerated statements and even remote control). Of course, some fairly basic constraints of business ethics and law apply to robots too: product safety and liability, or nondeception in advertisement. It appears that these existing constraints take care of many concerns that are raised. There are cases, however, in which HRI has aspects that appear specifically human in ways that can perhaps not be replaced by robots: care, love, and sex.

Example A: Care robots

The use of robots in health care for humans is currently at the level of concept studies in real environments, but it may become a usable technology in a few years and has raised a number of concerns for a dystopian future of dehumanized care (Sharkey and Sharkey 2011; Sparrow 2016). Current systems include robots that support human carers (caregivers)—for example, in lifting patients or transporting material; robots that enable patients to do certain things by themselves, such as eat with a robotic arm; and also robots that are given to patients as company and comfort (e.g., the "Paro" robot seal). For an overview, see (van Wynsberghe 2016; Fosch-Villaronga and Albo-Canals 2019; Nørskov 2017) and for a survey of users Draper et al. (2014).

One reason why the issue of care has come to the fore is that people have argued we will need robots in aging societies. This argument makes problematic assumptions—namely, that with longer life spans people will need more care and that it will not be possible to attract more humans to caring professions. It may also show a bias about age (Jecker 2020). Most importantly, it ignores the nature of automation, which is not simply about replacing humans but about allowing humans to work more effectively. It is not very clear that there really is an issue here since the discussion mostly focuses on the fear of robots dehumanizing care, but the actual and foreseeable robots in care are for the classic automation of technical tasks as assistive robots. They are thus "care robots" only in a behavioral sense of doing what is required, not in the sense that a human "cares" for the patients. It appears that the success of "being cared for" relies on this intentional sense of "care," which foreseeable robots cannot provide. If anything, the risk of robots in care is the *absence* of

such intentional care—because fewer human carers may be needed. Interestingly, caring for something, even a virtual agent, can be good for the carer themselves (Lee et al. 2019). A system that pretends to care would be deceptive and thus problematic—unless the deception is countered by sufficiently large utility gain (Coeckelbergh 2016). Some robots that pretend to "care" on a basic level are available (Paro seal), and others are in the making. Perhaps feeling cared for by a machine, to some extent, can be progress in some cases?

Example B: Sex robots

Several tech optimists have argued that humans will likely be interested in sex and companionship with robots and feel good about it (Levy 2007). Given the variation of human sexual preferences, including sex toys and sex dolls, this seems very likely: the question is whether such devices should be manufactured and promoted and whether there should be limits to use in this touchy area. It seems to have moved into the mainstream of "robot philosophy" in recent times (Sullins 2012; Danaher and McArthur 2017; Sharkey et al. 2017; Bendel 2018; Devlin 2018).

Humans have long had deep emotional attachments to objects, so perhaps companionship or even love with a predictable android is attractive, especially to people who struggle with actual humans and already prefer dogs, cats, a bird, a computer, or a Tamagotchi. Danaher (2019b) argues against Nyholm and Frank (2017) that this can be true friendship and is thus a valuable goal. It certainly looks like such friendship might increase overall utility, even if lacking in depth. In all these areas, there is an issue of deception since a robot cannot (at present) mean what it says or have feelings for a human. It is well known that humans are prone to attribute feelings and thoughts to entities that behave as if they had sentience and even to clearly inanimate objects that show no behavior at all. Also, paying for deception seems to be an elementary part of the traditional sex industry.

Finally, there are concerns that have often accompanied matters of sex—namely, consent (Frank and Nyholm 2017), aesthetic issues, and worry that humans may be "corrupted" by certain experiences. Old-fashioned though this may seem, human behavior is influenced by experience, and it is likely that pornography or sex robots support the perception of other humans as mere objects of desire, or even as recipients of abuse, and thus ruin a deeper sexual and erotic experience. The Campaign against Sex Robots argues that these devices are a continuation of slavery and prostitution (Richardson 2016).

12.2.2 The Effects of Automation on Employment

It seems clear that AI and robotics will lead to significant gains in productivity and thus overall wealth. The attempt to increase productivity has probably always been a feature of the economy, though the emphasis on "growth" is a modern phenomenon (Harari 2016, 240). However, productivity gains through automation typically mean that fewer humans are required for the same output. This does not necessarily imply a loss of overall employment, however, because available wealth increases and that can increase demand sufficiently to counteract the productivity gain. In the long run, higher productivity in industrial societies has led to more wealth overall. Major labor market disruptions have occurred in the past—for example, farming employed over 60 percent of the workforce in Europe and North America in 1800, while by 2010 it employed about 5 percent in the European Union

and even less in the wealthiest countries (Anonymous 2013). In the twenty years between 1950 and 1970, the number of hired agricultural workers in the UK was reduced by 50 percent (Zayed and Loft 2019). Some of these disruptions lead to more labor-intensive industries moving to places with lower labor cost—this is an ongoing process.

Classic automation replaces human muscle, whereas digital automation replaces human thought or information processing—and unlike physical machines, digital automation is very cheap to duplicate (Bostrom and Yudkovski 2014). It may thus mean a more radical change in the labor market. So the main question is: Is it different, this time? Will the creation of new jobs and wealth keep up with the destruction of jobs? And even if it is *not* different, what are the transition costs, and who bears them? For example, will lower-cost areas suffer and higher-cost areas gain from this development? Do we need to make societal adjustments for a fair distribution of costs and benefits of digital automation?

Responses to the issue of unemployment from robotics and AI have ranged from the alarmed (Frey and Osborne 2013; Westlake 2014) to the neutral (Metcalf, Keller, and Boyd 2016; Calo 2018; Frey 2019) and the optimistic (Brynjolfsson and McAfee 2016; Harari 2016; Danaher 2019a). In principle, the labor market effect of automation seems to be fairly well understood as involving two channels: "(i) the nature of interactions between differently skilled workers and new technologies affecting labor demand and (ii) the equilibrium effects of technological progress through consequent changes in labor supply and product markets" (Goos 2018, 362). What currently seems to happen in the labor market as a result of automation is "job polarization" or the "dumbbell" shape (Goos, Manning, and Salomons 2009): the highly skilled technical jobs are in demand and highly paid, the low-skilled service jobs are in demand and badly paid, but the midqualification jobs in factories and offices—that is, the majority of jobs—are under pressure and reduced because they are relatively predictable and most likely to be automated (Baldwin 2019).

Perhaps enormous productivity gains allow the "age of leisure" to be realized, which Keynes (1930) predicted to occur around 2030, assuming a growth rate of 1 percent per annum? Actually, we have already reached the level he anticipated for 2030, but we are still working—consuming more and inventing ever more levels of organization. Harari explained how this economical development allowed humanity to overcome hunger, disease, and war, and now we aim for immortality and eternal bliss through AI, thus his title *Homo Deus* (Harari 2016, 75).

In general terms, the issue of unemployment is one of how goods in a society should be *justly distributed*. A standard view is that distributive justice should be rationally decided from behind a "veil of ignorance" (Rawls 1971)—that is, as if one does not know what position in a society one would actually be taking (laborer or industrialist, and so on). Rawls thought the chosen principles would then support basic liberties and a distribution that is of greatest benefit to the least-advantaged members of society. It would appear that the robotics economy has three features that make such justice unlikely: First, it operates in a largely unregulated environment where responsibility is often hard to allocate. Second, it operates in markets that have a "winner-takes-all" feature, where monopolies develop quickly. Third, the "new economy" of the digital service industries is based on intangible assets, also called "capitalism without capital" (Haskel and Westlake 2017). This means that it is difficult to control multinational digital corporations that do not rely

on a physical plant in a particular location. These three features seem to suggest that if we leave the distribution of wealth to free market forces, the result would be a heavily unjust distribution. And this is indeed a development that we can already see.

One interesting question that has not received too much attention is whether the development of robotics is environmentally sustainable. Like all computing systems, they produce waste that is very hard to recycle, and they consume vast amounts of energy, especially for the training of machine-learning systems (and even for the mining of cryptocurrency). Again it appears that some agents off-load costs to the general society.

12.2.3 Privacy and Surveillance

There is a general discussion about privacy and surveillance in information technology (e.g., Macnish 2017; Roessler 2017), which mainly concerns the access to private data and data that are personally identifiable. Privacy has several well-recognized aspects—for example, "the right to be left alone," information privacy, privacy as an aspect of personhood, control over information about oneself, and the right to secrecy (Bennett and Raab 2006). Privacy studies have historically focused on state surveillance by secret services but now include surveillance by other state agents, businesses, and even individuals. The technology has changed massively in the last decades, while regulation has been slow to respond (though there is the GDPR [2016]). The result is an anarchy that is exploited by the most powerful players—sometimes in plain sight, sometimes in hiding.

The digital sphere has widened massively: all data collection and storage are now digital, our lives are more and more digital, most digital data are connected to a single internet, and there is more and more sensor technology around that generates data about nondigital aspects of our lives. At the same time, control over who collects which data, and who has access, is much harder in the digital world than it was in the analog world of paper and telephone calls. Every new technology amplifies the known issues. For example, face recognition in photos and videos allows identification and thus profiling and searching for individuals (Whittaker et al. 2018, 15ff). The result is that "in this vast ocean of data, there is a frighteningly complete picture of us" (Smolan 2016, 1:1), a scandal that still has not received due public attention.

The data trail we leave behind is how our "free" services are paid for, but we are not told about that data collection and its value, and we are manipulated into leaving ever more such data. The primary focus of social media, gaming, and most of the internet in this "surveillance economy" is to gain, maintain, and direct attention—and thus data supply. This surveillance and attention economy is sometimes called "surveillance capitalism" (Zuboff 2019).

Such systems will often reveal facts about us that we ourselves wish to suppress or are not aware of. With the last sentence of his best-selling book *Homo Deus*, Harari (2016) asks about the long-term consequences of AI: "What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?"

Robotic devices have not yet played a major role in this area, except for security patrolling, but this will change once they are more common outside of industry environments. Together with the Internet of Things, the "smart" systems (phone, TV, oven, lamp, virtual assistant, home . . .), the "smart city" (Sennett 2018), and "smart governance," they are set to become part of the data-gathering machinery that offers more detailed data, of different types, in real time, with ever more information.

Privacy-preserving techniques that can conceal the identity of persons or groups to a large extent are now a standard staple in data science; they include (relative) anonymization, access control (plus encryption), and other models in which computation is carried out without access to full unencrypted input data (Stahl and Wright 2018), in the case of "differential privacy" by adding calibrated noise to the output of queries (Dwork et al. 2006; Abowd 2017). While requiring more effort and cost, such techniques can avoid many of the privacy issues. Some companies have also seen better privacy as a competitive advantage that can be leveraged and sold at a price.

12.2.4 Autonomous Systems

Autonomy generally

Several notions of autonomy can be found in the discussion of autonomous systems. A stronger notion is involved in philosophical debates in which autonomy is the basis for responsibility and personhood (Christman 2018). In this context, responsibility implies autonomy, but not inversely, so some systems can have degrees of technical autonomy without raising issues of responsibility. The weaker, more technical, notion of autonomy in robotics is relative and gradual: a system is said to be autonomous with respect to human control to a certain degree (Müller 2012). There is a parallel here to the issues of bias and opacity in AI since autonomy also concerns a power relation: Who is in control, and who is responsible?

Generally speaking, one question is whether autonomous robots raise issues that suggest a revision of present conceptual schemes or whether they just require technical adjustments. In most jurisdictions, there is a sophisticated system of civil and criminal liability to resolve such issues. Technical standards—for example, for the safe use of machinery in medical environments—will likely need to be adjusted. There is already a field of "verifiable AI" for such safety-critical systems and for "security applications." Bodies like the IEEE and the BSI have produced "standards," particularly for more technical subproblems, such as data security and transparency. Among the many autonomous systems on land, on water, underwater, in the air, or in space, we discuss two samples: autonomous vehicles and autonomous weapons.

Example A: Autonomous vehicles

Autonomous vehicles hold the promise of reducing the very significant damage that human driving currently causes—with approximately one million humans killed per year, many more injured, the environment polluted, the earth sealed with concrete and tarmac, the cities full of parked cars, and so on. However, there seem to be questions of how autonomous vehicles should behave and how responsibility and risk should be distributed in the complicated system the vehicles operate in. (There is also significant disagreement over how long the development of fully autonomous, or "level 5," cars [SAE 2015] will actually take.)

There is some discussion of "trolley problems" in this context. In the classic trolley problems (Thompson 1976; Woollard and Howard-Snyder 2016, sect. 2), various dilemmas are presented. The simplest version is that of a trolley train on a track that is heading toward five people and will kill them unless the train is diverted onto a side track. However, on that track is one person who will be killed if the train takes that side track. The example goes back to a remark in (Foot 1967, 6), who discusses a number of dilemma cases in which tolerated and intended consequences of an action differ. Trolley problems are not

supposed to describe actual ethical problems or to be solved with a "right" choice. Rather, they are thought experiments in which choice is artificially constrained to a small, finite number of distinct one-off options and where the agent has perfect knowledge. These problems are used as a theoretical tool to investigate ethical intuitions and theories—especially the difference between actively doing versus allowing something to happen, intended versus tolerated consequences, and consequentialist versus other normative approaches (Kamm and Rakowski 2016). This type of problem has reminded many of the problems encountered in actual driving and in autonomous driving (Lin 2015). It is doubtful, however, that an actual driver or autonomous car will ever have to solve trolley problems (but see Keeling 2019). While autonomous car trolley problems have received a lot of media attention (Awad et al. 2018), they do not seem to offer anything new to either ethical theory or to the programming of autonomous vehicles.

The more common ethical problems in driving, such as speeding, risky overtaking, not keeping a safe distance, and more are classic problems of pursuing personal interest versus the common good. The vast majority of these are covered by legal regulations on driving. Programming the car to drive "by the rules" rather than "by the interest of the passengers" or "to achieve maximum utility" is thus deflated to a standard problem of programming ethical machines (see section 3.1). There are probably additional discretionary rules of politeness and interesting questions on when to break the rules (Lin 2015), but again this seems to be more a case of applying standard considerations (rules vs. utility) to autonomous vehicles.

Notable policy efforts in this field include the report by the German Federal Ministry of Transport and Digital Infrastructure (2017), which stresses that *safety* is the primary objective. Rule 10 states, "In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions" (see 3.2.1). The resulting German and EU laws on licensing automated driving are much more restrictive than their US counterparts, where "testing on consumers" is a strategy used by some companies—without informed consent of the consumers or the possible victims.

Example B: Autonomous weapons

The notion of automated weapons is fairly old: "For example, instead of fielding simple guided missiles or remotely piloted vehicles, we might launch completely autonomous land, sea, and air vehicles capable of complex, far-ranging reconnaissance and attack missions" (DARPA 1983, 1). This proposal was ridiculed as "fantasy" at the time (Dreyfus, Dreyfus, and Athanasiou 1986, ix), but it is now a reality, at least for more easily identifiable targets (missiles, planes, ships, tanks, and so on) but not for human combatants. The main arguments against (lethal) autonomous weapon systems (AWS or LAWS) are that they support extrajudicial killings, take responsibility away from humans, and make wars or killings more likely—for a detailed list of issues see (Lin, Bekey, and Abney 2008, 73–86).

It appears that lowering the hurdle to use such systems (autonomous vehicles, "fire-and-forget" missiles, or drones loaded with explosives) and reducing the probability of being held accountable would increase the probability of their use. The crucial asymmetry in which one side can kill with impunity and thus has few reasons not to do so already

exists in conventional drone wars with remote-controlled weapons (e.g., the US in Pakistan). It is easy to imagine a small drone that searches, identifies, and kills an individual human—or perhaps a type of human. These are the kinds of cases brought forward by the Campaign to Stop Killer Robots and other activist groups. Some seem to be equivalent to saying that autonomous weapons are indeed weapons, and weapons kill, but we still make them in gigantic numbers. On the matter of accountability, autonomous weapons might make the identification and prosecution of the responsible agents more difficult, but this is not clear given the digital records that one can keep, at least in a conventional war. The difficulty of allocating punishment is sometimes called the "retribution gap" (Danaher 2016).

Another question seems to be whether using autonomous weapons in war would make wars worse or perhaps less bad? If robots reduce war crimes and crimes in war, the answer may well be positive and has been used not only as an argument in favor of these weapons (Arkin 2009; Müller 2016) but also as an argument against (Amoroso and Tamburrini 2018). Arguably, the main threat is not the use of such weapons in conventional warfare but in asymmetric conflicts or by nonstate agents, including criminals.

It has also been said that autonomous weapons cannot conform to International Humanitarian Law, which requires observance of the principles of distinction (between combatants and civilians), proportionality (of force), and military necessity (of force) in military conflict (Sharkey 2019). It is true that the distinction between combatants and noncombatants is difficult to discern, but the distinction between civilian and military ships is easy to see—so all this says is that we should not construct and use such weapons if they do violate humanitarian law. Additional concerns have been raised that being killed by an autonomous weapon threatens human dignity, but even the defenders of a ban on these weapons seem to say that these are not good arguments: "There are other weapons, and other technologies, that also compromise human dignity. Given this, and the ambiguities inherent in the concept, it is wiser to draw on several types of objections in arguments against AWS, and not to rely exclusively on human dignity" (Sharkey 2019).

A lot has been made of keeping humans "in the loop" or "on the loop" of military guidance on weapons—these ways of spelling out "meaningful control" are discussed in Santoni de Sio and van den Hoven (2018). There have been discussions about the difficulties of allocating responsibility for the killings of an autonomous weapon, and a "responsibility gap" has been suggested (esp. Sparrow 2007), meaning that neither the human nor the machine may be responsible. On the other hand, we do not assume that for every event there is someone responsible for that event, and the real issue may well be the distribution of risk (Simpson and Müller 2016). Risk analysis (Hansson 2013) indicates it is crucial to identify who is *exposed* to risk, who is a potential *beneficiary*, and who makes the *decisions* (Hansson 2018, 1822–1824).

12.3 Ethics for Robotics Systems

12.3.1 Machine Ethics

Machine ethics is ethics for machines, for "ethical machines," and for machines as *subjects* rather than for the human use of machines as *objects*. It is often not very clear whether this is supposed to cover all of robot ethics of to be a part of it (Floridi and Saunders 2004;

Moor 2006; Wallach and Asaro 2017; Anderson and Anderson 2011). Sometimes it looks as though there is the (dubious) inference at play here that if machines act in ethically relevant ways, then we need a machine ethics. Accordingly, some use a broader notion: "Machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable" (Anderson and Anderson 2007, 15). This might include mere matters of product safety, for example. Some of the discussion in machine ethics makes the very substantial assumption that machines can, in some sense, be ethical agents responsible for their actions, or "autonomous moral agents" (see van Wynsberghe and Robbins 2019). The basic idea of machine ethics is now finding its way into actual robotics, where the assumption that these machines are artificial moral agents in any substantial sense is usually not made (Winfield et al. 2019). It is sometimes observed that a robot that is programmed to follow ethical rules can very easily be modified to follow unethical rules (Vanderelst and Winfield 2018).

The idea that machine ethics might take the form of "laws" has famously been investigated by Isaac Asimov (1942), who proposed "three laws of robotics": "First Law—A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law—A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law—A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws." Asimov then showed in a number of stories how conflicts between these three laws will make it problematic to use them, despite their hierarchical organization.

It is not clear that there is a consistent notion of "machine ethics" since weaker versions are in danger of reducing "having an ethics" to notions that would not normally be considered sufficient (e.g., without "reflection" or even without "action"); stronger notions that move toward artificial moral agents may describe a—currently—empty set.

12.3.2 Artificial Moral Agents

If one takes machine ethics to concern moral agents, in some substantial sense, then these agents can be called "artificial moral agents" having rights and responsibilities. However, the discussion about artificial entities challenges a number of common notions in ethics, and it can be very useful to understand these in abstraction from the human case (cf. Powers and Ganascia, forthcoming; Misselhorn 2020).

Several authors use "artificial moral agent" in a less demanding sense, borrowing from the software "agent" use in which case matters of responsibility and rights will not arise (Allen, Varner, and Zinser 2000). James Moor (2006) distinguishes four types of machine agents: ethical impact agents (example: robot jockeys), implicit ethical agents (example: safe autopilot), explicit ethical agents (example: using formal methods to estimate utility), and full ethical agents ("Can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent"). Several ways to achieve "explicit" or "full" ethical agents have been proposed, via programming it in (operational morality), via "developing" the ethics itself (functional morality), and finally, full-blown morality with full intelligence and sentience (Allen, Smit, and Wallach 2005; Moor 2006). Programmed agents are sometimes not considered "full" agents because they are "competent without comprehension," just like the neurons in a brain (Dennett 2017; Hakli and Mäkelä 2019).

In some of these discussions, the notion of "moral patient" plays a role: ethical *agents* have responsibilities, while ethical *patients* have rights, because harm to them matters. It seems clear that some entities are patients without being agents—for example, simple animals that can feel pain but cannot make justified choices. On the other hand, it is normally understood that all agents will also be patients (e.g., in a Kantian framework). Usually, being a person is supposed to be what makes an entity a responsible agent, someone who can have duties and be the object of ethical concerns, and such personhood is typically a deep notion associated with free will (Frankfurt 1971; Strawson 2005) and with having phenomenal consciousness. Torrance (2011) suggests "artificial (or machine) ethics could be defined as designing machines that do things which, when done by humans, are criterial of the possession of 'ethical status' in those humans"—which he takes to be "ethical *productivity* and ethical *receptivity*"—his expressions for moral agents and patients.

Responsibility for robots

There is broad consensus that accountability, liability, and the rule of law are basic requirements that must be upheld in the face of new technologies (European Group on Ethics in Science and New Technologies 2018, 18), but the issue is how this can this be done and how responsibility can be allocated. If the robots act, will they themselves be responsible, liable, or accountable for their actions? Or should the distribution of risk perhaps take precedence over discussions of responsibility?

Traditional distribution of responsibility already occurs: a car maker is responsible for the technical safety of the car, a driver is responsible for driving, a mechanic is responsible for proper maintenance, the public authorities are responsible for the technical conditions of the roads, and so on. In general "the effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware. . . . With distributed agency comes distributed responsibility" (Taddeo and Floridi 2018, 751). How this distribution might occur is not a problem that is specific to robotics, but it gains particular urgency in this context (Nyholm 2018a, 2018b).

Rights for robots

Some authors have indicated that whether or not current robots must be allocated rights should be seriously considered (Gunkel 2018a, 2018b; Turner 2019; Danaher 2020). This position seems to rely largely on criticism of the opponents and on the empirical observation that robots and other nonpersons are sometimes treated as having rights. In this vein, a "relational turn" has been proposed: If we relate to robots as though they had rights, then we might be well advised not to search whether they "really" do have such rights, but we should assume that they do (Coeckelbergh 2010, 2012, 2018). This raises the question of how far such antirealism or quasi-realism can go and what it means then to say that "robots have rights" in a human-centered approach (Gerdes 2016). On the other side of the debate, Bryson (2010) has insisted with a useful (but admittedly problematic) slogan that "robots should be slaves"—that is, not enjoy rights, though she considers it a possibility (Gunkel and Bryson 2014).

There is a wholly separate issue of whether robots (or other AI systems) should be given the status of "legal entities" or "legal persons"—in the sense in which natural persons but also states, businesses, or organizations are "entities" and can have legal rights and duties. The European Parliament has considered allocating such status to robots in order to deal

with civil liability (EU Parliament 2016; Bertolini and Aiello 2018) but not criminal liability, which is reserved for natural persons. It would also be possible to assign only a certain subset of rights and duties to robots. It has been said that "such legislative action would be morally unnecessary and legally troublesome" because it would not serve the interest of humans (Bryson, Diamantis, and Grant 2017, 273). In environmental ethics there is a long-standing discussion about the legal rights for natural objects like trees (Stone 1972).

It has also been said that the reasons for developing robots with rights, or artificial moral patients, in the future are ethically doubtful (van Wynsberghe and Robbins 2019). In the community of "artificial consciousness" researchers is significant concern about whether it would be ethical to create such consciousness since this would presumably imply ethical obligations to a sentient being—for example, not to harm it and not to end its existence by switching it off. Some authors have called for a "moratorium on synthetic phenomenology" (Bentley et al. 2018, 28f).

12.4 Conclusion

It is remarkable how imagination or a "vision" of robotics and AI has played a central role since the very beginning of the disciplines in the 1950s. And the evaluation of this vision is subject to dramatic change: In a few decades, we went from the slogans "AI is impossible" (Dreyfus) and "AI is just automation" (Lighthill 1973) to "AI will solve all problems" (Kurzweil 1999) and "AI may kill us all" (Bostrom 2014). This created media attention and public relations efforts, but it also raises the problem of how much of this "philosophy and ethics of AI and robotics" is really an imagined technology. As we said at the outset, AI and robotics have raised fundamental questions about what we should do with these systems, what the systems themselves should do, and what risks they have in the long term. They also challenge the human view of humanity as the intelligent and dominant species on Earth. We have seen the issues that have been raised, and we will have to watch technological and social developments closely to catch the new issues early and to develop a philosophical analysis, as well as to debate the traditional problems of philosophy.

Acknowledgments

This chapter has significant overlap with the article by the same author: Müller, Vincent C. 2020. "Ethics of Artificial Intelligence and Robotics." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, 1–70. Palo Alto: CSLI, Stanford University. https://plato.stanford.edu/entries/ethics-ai/—I am grateful for the comments of many colleagues on that version.

Parts of the work on this article have been supported by the European Commission under the INBOTS project (H2020 grant no. 780073).

Additional Reading and Resources

- Classic book arguing for the existential risk from AI: Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

• Short and classic introduction to machine ethics: Moor, James H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21.

• Textbook on robot ethics: Royakkers, Lambèr, and Rinie van Est. 2016. *Just Ordinary Robots: Automation from Love to War*. Boca Raton: CRC Press; Taylor and Francis.

• Newsletter on AI ethics in Europe (Charlotte Stix): https://www.charlottestix.com/europeanaiarchive.

References

Abowd, John M. 2017. "How Will Statistical Agencies Operate When All Data Are Private?" *Journal of Privacy and Confidentiality* 7 (3): 1–15.

AI HLEG. 2019. "High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI." European Commission. Last modified March 8, 2021. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches." *Ethics and Information Technology* 7 (3): 149–155.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12 (3): 251–261.

Amoroso, Daniele, and Guglielmo Tamburrini. 2018. "The Ethical and Legal Case against Autonomy in Weapons Systems." *Global Jurist* 18 (1).

Anderson, Michael, and Susan Leigh Anderson. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28 (4): 15–26.

Anderson, Michael, and Susan Leigh Anderson, eds. 2011. *Machine Ethics*. Cambridge: Cambridge University Press.

Anonymous. 2013. "How Many People Work in Agriculture in the European Union? An Answer Based on Eurostat Data Sources." *EU Agricultural Economics Briefs* 8.

Arkin, Ronald C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.

Arnold, Thomas, and Matthias Scheutz. 2017. "Beyond Moral Dilemmas: Exploring the Ethical Landscape in Hri." In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction*, 445–452. New York: IEEE.

Asimov, Isaac. 1942 [1950]. "Runaround: A Short Story." *Astounding Science Fiction*. Reprinted in *I, Robot*. New York: Gnome Press, 1950, 40ff.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. "The Moral Machine Experiment." *Nature* 563 (7729): 59–64.

Baldwin, Richard. 2019. *The Globotics Upheaval: Globalisation, Robotics and the Future of Work*. London: Weidenfeld and Nicolson.

Bendel, Oliver. 2018. "Sexroboter aus Sicht der Maschinenethik." In *Handbuch Maschinenethik*, edited by Oliver Bendel, 1–19. Wiesbaden: Springer Fachmedien Wiesbaden.

Bennett, Colin J., and Charles Raab. 2006. *The Governance of Privacy: Policy Instruments in Global Perspective*. 2nd ed. Cambridge, MA: MIT Press.

Bentley, Peter J., Miles Brundage, Olle Häggström, and Thomas Metzinger. 2018. "Should We Fear Artificial Intelligence? In-Depth Analysis." *European Parliamentary Research Service, Scientific Foresight Unit* 614 (547): 1–40.

Bertolini, Andrea, and Giuseppe Aiello. 2018. "Robot Companions: A Legal and Ethical Analysis." *Information Society* 34 (3): 130–140.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, Nick, and Eliezer Yudkovski. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish, 316–334. Cambridge: Cambridge University Press.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." ArXiv preprint: 1802.07228.

Brynjolfsson, Erik, and Andrew McAfee. 2016. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton.

Bryson, Joanna J. 2010. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks, 63–74. Amsterdam: John Benjamins.

Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. 2017. "Of, For, and By the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25 (3): 273–291.

Calo, Ryan. 2018. "Artificial Intelligence Policy: A Primer and Roadmap." *University of Bologna Law Review* 3 (2): 180–218.

Calo, Ryan, Michael A. Froomkin, and Ian Kerr, eds. 2016. *Robot Law*. Cheltenham: Edward Elgar.

Christman, John. 2018. "Autonomy in Moral and Political Philosophy." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Palo Alto: Stanford University.

Coeckelbergh, Mark. 2010. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." *Ethics and Information Technology* 12 (3): 209–221.

Coeckelbergh, Mark. 2012. *Growing Moral Relations: Critique of Moral Status Ascription*. London: Palgrave.

Coeckelbergh, Mark. 2016. "Care Robots and the Future of ICT-Mediated Elderly Care: A Response to Doom Scenarios." *AI and Society* 31 (4): 455–462.

Coeckelbergh, Mark. 2018. "What Do We Mean by a Relational Ethics? Growing a Relational Approach to the Moral Standing of Plants, Robots and Other Non-humans." In *Plant Ethics*, edited by Angela Kallhoff, Marcello Di Paola, and Maria Schörgenhumer, 110–121. London: Routledge.

Crawford, Kate, and Ryan Calo. 2016. "There Is a Blind Spot in AI Research." *Nature* 538 (7625): 311–313.

Danaher, John. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18 (4): 299–309.

Danaher, John. 2019a. *Automation and Utopia: Human Flourishing in a World without Work*. Cambridge, MA: Harvard University Press.

Danaher, John. 2019b. "The Philosophical Case for Robot Friendship." *Journal of Posthuman Studies* 3 (1): 5–24.

Danaher, John. 2020. "Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviorism." *Science and Engineering Ethics* 26: 2023–2049.

Danaher, John, and Neil McArthur, eds. 2017. *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.

DARPA (Defense Advanced Research Projects Agency). 1983. *Strategic Computing: New Generation Computing Technology, a Strategic Plan for Its Development and Application to Critical Problems in Defense*. October 28, 1983. https://www.nitrd.gov/nitrdgroups/images/3/3a/20040929_strategic_computing.pdf.

Dennett, Daniel C. 2017. *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W. W. Norton.

Devlin, Kate. 2018. *Turned On: Science, Sex and Robots*. London: Bloomsbury.

Draper, Heather, Tom Sorell, Sandra Bedaf, Dag Sverre Syrdal, Carolina Gutierrez-Ruiz, Alexandre Duclos, and Farshid Amirabdollahian. 2014. "Ethical Dimensions of Human-Robot Interactions in the Care of Older People: Insights from 21 Focus Groups Convened in the UK, France and the Netherlands." In *International Conference on Social Robotics*, edited by M. Beetz, B. Johnston, and M. A. Williams. Cham, Switzerland: Springer.

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. 2nd ed. Cambridge, MA: MIT Press.

Dreyfus, Hubert L., Stuart E. Dreyfus, and Tom Athanasiou. 1986. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*. Berlin: Springer.

EU Parliament. 2016. *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. January 27, 2017. https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html.

European Group on Ethics in Science and New Technologies. 2018. *Statement on Artificial Intelligence, Robotics and "Autonomous" Systems*. Last modified September 3, 2018. http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

Floridi, Luciano, and Jeff W. Saunders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14:349–379.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect." *Oxford Review* 5:5–15.

Fosch-Villaronga, Eduard, and Jordi Albo-Canals. 2019. "'I'll Take Care of You,' Said the Robot: Reflecting upon the Legal and Ethical Aspects of the Use and Development of Social Robots for Therapy." *Paladyn, Journal of Behavioral Robotics* 10 (1): 77–93.

Frank, Lily, and Sven Nyholm. 2017. "Robot Sex and Consent: Is Consent to Sex between a Robot and a Human Conceivable, Possible, and Desirable?" *Artificial Intelligence and Law* 25 (3): 305–323.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5–20.

Frey, Carl Benedict. 2019. *The Technology Trap: Capital, Labour, and Power in the Age of Automation*. Princeton, NJ: Princeton University Press.

Frey, Carl Benedict, and Michael A. Osborne. 2013. "The Future of Employment: How Susceptible are Jobs to Computerisation?" Oxford Martin School Working Papers. September 1, 2013. https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/.

GDPR. 2016. "General Data Protection Regulation: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC." *Official Journal of the European Union* 119:1–88.

Gerdes, Anne. 2016. "The Issue of Moral Consideration in Robot Ethics." *SIGCAS Computers and Society* 45 (3): 274–279.

German Federal Ministry of Transport and Digital Infrastructure. 2017. *Report of the Ethics Commission: Automated and Connected Driving*. Federal Ministry of Transport and Digital Infrastructure. June 2017. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

Giubilini, Alberto, and Julian Savulescu. 2018. "The Artificial Moral Advisor: The 'Ideal Observer' Meets Artificial Intelligence." *Philosophy and Technology* 31 (2): 169–188.

Goos, Maarten. 2018. "The Impact of Technological Progress on Labour Markets: Policy Challenges." *Oxford Review of Economic Policy* 34 (3): 362–375.

Goos, Maarten, Alan Manning, and Anna Salomons. 2009. "Job Polarization in Europe." *American Economic Review* 99 (2): 58–63.

Gunkel, David J. 2018a. "The Other Question: Can and Should Robots Have Rights?" *Ethics and Information Technology* 20 (2): 87–99.

Gunkel, David J. 2018b. *Robot Rights*. Cambridge, MA: MIT Press.

Gunkel, David J., and Joanna Bryson. 2014. "Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient." *Philosophy and Technology* 27(1): 5–8.

Hakli, Raul, and Pekka Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *Monist* 102 (2): 259–275.

Hansson, Sven Ove. 2013. *The Ethics of Risk: Ethical Analysis in an Uncertain World*. New York: Palgrave Macmillan.

Hansson, Sven Ove. 2018. "How to Perform an Ethical Risk Analysis (Era)." *Risk Analysis* 38 (9): 1820–1829.

Harari, Yuval Noah. 2016. *Homo Deus: A Brief History of Tomorrow*. New York: Harper.

Haskel, Jonathan, and Stian Westlake. 2017. *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton, NJ: Princeton University Press.

Houkes, Wybo, and Pieter E. Vermaas. 2010. *Technical Functions: On the Use and Design of Artefacts*. Berlin: Springer.

Jacobs, An, Lynn Tytgat, Michel Maus, Romain Meeusen, and Bram Vanderborght, eds. 2019. *Homo Roboticus: 30 Questions and Answers on Man, Technology, Science and Art*. Brussels: ASP.

Jasanoff, Sheila. 2016. *The Ethics of Invention: Technology and the Human Future*. New York: W. W. Norton.

Jecker, Nancy S. 2020. *Ending Midlife Bias: New Values for Old Age*. New York: Oxford University Press.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–399.

Johnson, Deborah G., and Mario Verdicchio. 2017. "Reframing AI Discourse." *Minds and Machines* 27 (4): 575–590.

Kamm, Frances Myrna, and Eric Rakowski, eds. 2016. *The Trolley Problem Mysteries*. New York: Oxford University Press.

Keeling, Geoff. 2019. "Why Trolley Problems Matter for the Ethics of Automated Vehicles." *Science and Engineering Ethics* 26 (1): 293–307.

Keynes, John Maynard. 1932. "Economic Possibilities for Our Grandchildren." In *Essays in Persuasion*, 358–373. New York: Harcourt Brace.

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. London: Penguin.

Lee, Minha, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand Ijsselsteijn. 2019. "Caring for Vincent: A Chatbot for Self-Compassion." *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, no. 702. https://doi.org/10.1145/3290605.3300932.

Levy, David. 2007. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper.

Lighthill, James. 1973. "Artificial Intelligence: A General Survey." In *Artificial Intelligence: A Paper Symposium*, 1–21. London: Science Research Council.

Lin, Patrick. 2015. "Why Ethics Matters for Autonomous Cars." In *Autonomous Driving*, edited by M. Maurer et al., 69–85. Berlin: Springer.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. 2017. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.

Lin, Patrick, George Bekey, and Keith Abney. 2008. "Autonomous Military Robotics: Risk, Ethics, and Design." US Department of Navy, Office of Naval Research. http://ethics.calpoly.edu/onr_report.pdf.

Macnish, Kevin. 2017. *The Ethics of Surveillance: An Introduction*. London: Routledge.

Metcalf, Jacob, Emily F. Keller, and Danah Boyd. 2016. "Perspectives on Big Data, Ethics, and Society." Council for Big Data, Ethics, and Society. May 23, 2016. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.

Misselhorn, Catrin. 2020. "Artificial Systems with Moral Capacities? A Research Design and Its Implementation in a Geriatric Care System." *Artificial Intelligence* 278:103179.

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

Müller, Vincent C. 2012. "Autonomous Cognitive Systems in Real-World Environments: Less Control, More Flexibility and Better Interaction." *Cognitive Computation* 4 (3): 212–215.

Müller, Vincent C. 2016. "Autonomous Killer Robots Are Probably Good News." In *Drones and Responsibility: Legal, Philosophical and Socio-technical Perspectives on the Use of Remotely Controlled Weapons*, edited By Ezio Di Nucci and Filippo Santoni De Sio, 67–81. London: Ashgate.

Nørskov, Marco, ed. 2017. *Social Robots*. London: Routledge.

Nyholm, Sven. 2018a. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24 (4): 1201–1219.

Nyholm, Sven. 2018b. "The Ethics of Crashes with Self-Driving Cars: A Roadmap, II." *Philosophy Compass* 13 (7): e12506.

Nyholm, Sven, and Lily Frank. 2017. "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?" In *Robot Sex: Social and Ethical Implications*, edited by John Danaher and Neil McArthur, 219–243. Cambridge, MA: MIT Press.

Powers, Thomas M., and Jean-Gabriel Ganascia. Forthcoming. "The Ethics of the Ethics of AI." In *Oxford Handbook of Ethics of Artificial Intelligence*, edited by Markus D. Dubber, Frank Pasquale, and Sunnit Das.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press.

Richardson, Kathleen. 2016. "Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines." *IEEE Technology and Society* 35 (2).

Roessler, Beate. 2017. "Privacy as a Human Right." *Proceedings of the Aristotelian Society* 2 (117).

Royakkers, Lambèr, and Rinie van Est. 2016. *Just Ordinary Robots: Automation from Love to War*. Boca Raton: CRC Press, Taylor and Francis.

SAE International. 2015. "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles." *SAE Recommended Practice* J3016_201806.

Santoni De Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5 (15).

Sennett, Richard. 2018. *Building and Dwelling: Ethics for the City*. London: Allen Lane.

Sharkey, Amanda. 2019. "Autonomous Weapons Systems, Killer Robots and Human Dignity." *Ethics and Information Technology* 21 (2): 75–87.

Sharkey, Amanda, and Noel Sharkey. 2011. "The Rights and Wrongs of Robot Care." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and George Bekey, 267–282. Cambridge, MA: MIT Press.

Sharkey, Noel, Aimee Van Wynsberghe, Scott Robbins, and Eleanor Hancock. 2017. "Report: Our Sexual Future with Robots." Responsible Robotics. July 5, 2017. https://responsiblerobotics.org/2017/07/05/frr-report-our-sexual-future-with-robots/.

Simpson, Thomas W., and Vincent C. Müller. 2016. "Just War and Robots Killings." *Philosophical Quarterly* 66 (263): 302–322.

Smolan, Sandy. 2016. "The Human Face of Big Data." PBS documentary. 56 mins.

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

Sparrow, Rob. 2016. "Robots in Aged Care: A Dystopian Future." *AI and Society* 31 (4): 1–10.

Stahl, Bernd Carsten, Job Timmermans, and Brent Daniel Mittelstadt. 2016. "The Ethics of Computing: A Survey of the Computing-Oriented Literature." *ACM Computing Surveys* 48/4 (55): 1–38.

Stahl, Bernd Carsten, and David Wright. 2018. "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation." *IEEE Security and Privacy* 16 (3).

Stone, Christopher D. 1972. "Should Trees Have Standing—toward Legal Rights for Natural Objects." *Southern California Law Review* 2:450–501.

Strawson, Galen. 2005. *Free Will*. London: Routledge. Last modified February 29, 2004. http://www.rep.routledge.com/article/v014.

Sullins, John P. 2012. "Robots, Love, and Sex: The Ethics of Building a Love Machine." *IEEE Transactions on Affective Computing* 3 (4): 398–409.

Taddeo, Mariarosaria, and Luciano Floridi. 2018. "How AI Can Be a Force for Good." *Science* 361 (6404): 751–752.

Thompson, Judith Jarvis. 1976. "Killing, Letting Die and the Trolley Problem." *Monist* 59:204–217.

Torrance, Steve. 2011. "Machine Ethics and the Idea of a More-than-Human Moral World." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 115–137. Cambridge: Cambridge University Press.

Turner, Jacob. 2019. *Robot Rules: Regulating Artificial Intelligence*. Berlin: Springer.

Tzafestas, Spyros G. 2016. *Roboethics: A Navigating Overview*. Berlin: Springer.

Vanderelst, Dieter, and Alan Winfield. 2018. "The Dark Side of Ethical Robots." In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 317–322. https://doi.org/10.1145/3278721.3278726.

van Wynsberghe, Aimee. 2016. *Healthcare Robots: Ethics, Design and Implementation*. London: Routledge.

van Wynsberghe, Aimee, and Scott Robbins. 2019. "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics* 25 (3): 719–735.

Verbeek, Peter-Paul. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.

Wallach, Wendell, and Peter M. Asaro, eds. 2017. *Machine Ethics and Robot Ethics*. London: Routledge.

Westlake, Stian, ed. 2014. *Our Work Here Is Done: Visions of a Robot Economy*. London: Nesta.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, and Jason Schultz. 2018. "AI Now Report 2018." New York University. https://ainowinstitute.org/ai_now_2018_report.html.

Winfield, Alan F., Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems." *Proceedings of the IEEE* 107 (3): 509–517.

Woollard, Fiona, and Frances Howard-Snyder. 2016. "Doing vs. Allowing Harm." In *Stanford Encyclopedia of Philosophy* Fall 2021 edition, edited by Edward N. Zalta. Palo Alto: Stanford University. https://plato.stanford.edu/archives/fall2021/entries/doing-allowing/.

Zayed, Yago, and Philip Loft. 2019. "Agriculture: Historical Statistics." *House of Commons Briefing Paper* 3339:1–19.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

This is a section of doi:10.7551/mitpress/13780.001.0001

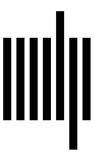
Cognitive Robotics

Edited by: Angelo Cangelosi, Minoru Asada

Citation:

Cognitive Robotics
Edited by: Angelo Cangelosi, Minoru Asada
DOI: 10.7551/mitpress/13780.001.0001
ISBN (electronic): 9780262369329
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Times New Roman by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Cangelosi, Angelo, 1967– editor. | Asada, Minoru, editor.
Title: Cognitive robotics / edited by Angelo Cangelosi and Minoru Asada.
Other titles: Cognitive robotics (M.I.T. Press)
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Series: Intelligent robotics and autonomous agents series | Includes bibliographical references and index.
Identifiers: LCCN 2021031320 | ISBN 9780262046831 (hardcover)
Subjects: LCSH: Autonomous robots.
Classification: LCC TJ211.35 .C628 2022 | DDC 629.8/92—dc23
LC record available at https://lccn.loc.gov/2021031320