The Ethical Landscape of Robotics

Bringing Ethics into the Design and Use of Robots

By Pawel Łichocki, Peter H. Kahn, Jr., and Aude Billard

This decade has undergone a true robotic demographic explosion. The number of industrial robots in operation exceeded 1 million by the end of 2008. Sales of robots for personal and domestic purposes have increased significantly since 2000 and reached 7.2 million by the end of 2009 [41]. The rampant growth of service robots led to rethink about the role of robots within the human society. Robots are no longer slave machines that respond purely to human requests. They are warranted for some degree of autonomy and decision making. Some, even, envision-friendly and entertaining robots that may become our companions. As a result of this recent robot emancipation, a number of ethical issues have emerged that were not relevant before. We believe that a lively and engaged discussion of ethical issues in robotics by roboticists and others is essential for creating a better and more just world.

In this article, we highlight the possible benefits, as well potential threats, related to the widespread use of robots. We follow the view that a robot cannot be analyzed on its own without taking into consideration the complex sociotechnical nexus of today's societies and that high-tech devices, such as robots, may influence how societies develop in ways that could not be foreseen during the design of the robots. In our survey, we limit ourselves to presenting the ethical issues delineated by other authors and relay their lines of reasoning for raising the public's concerns. We show that disagreements on what is ethical or not in robotics stem often from different beliefs on human nature and different expectations on what technology may achieve in the future. We do not offer a personal stance to these issues, so as to allow the reader to form his/her opinion.

In terms of robotic applications, we focus on service robots that peacefully interact with humans [Figure 1(a) and (b)] and lethal robots created to fight on battlefields [Figure 1(c) and (d)]. Other robotic applications are also discussed in the literature; therefore, various concerns for our societies are not discussed here. Unfortunately, for space constraints, we had to limit ourselves in our presentation. For instance, we omitted the question of unemployment caused by the development of industrial robots. This concern is in line with the general issue of using machines to replace human labor, a topic that is central to philosophical debates since the industrial revolution. Furthermore, we chose not to discuss the concerns that robots may one day be able to claim some social, cultural, ethical, or legal rights, that

Digital Object Identifier 10.1109/MRA.2011.940275
Date of publication: 14 April 2011

1070-9932/11/$26.00©2011 IEEE

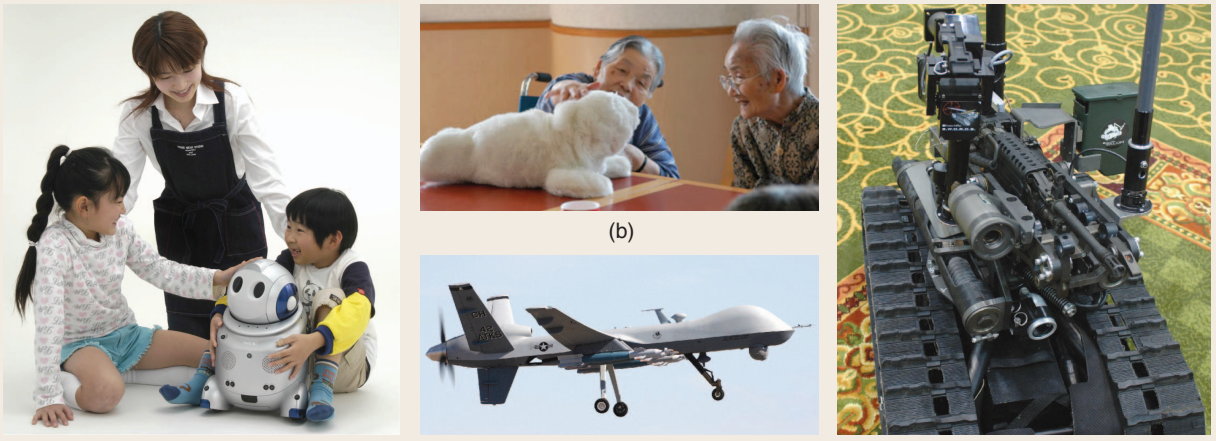

Figure 1. Robotic applications of (a), (b) service and (c), (d) combat robots. (a) Childcare robot PaPeRo [32], [73]. [Photo courtesy of NEC Corporation.] (b) Paro therapeutic robot [89]. [Photo courtesy of AIST, Japan.] (c) MQ-9 Reaper Hunter/Killer UAV by General Atomics Aeronautical Systems [33]. (d) Special weapons observation reconnaissance detection system (SWORDS) by Foster-Miller [42]. [Photo courtesy of Foster-Miller.]

robots may become sentient machines [51], which we would no longer be allowed to enslave [75], or that we may create robots capable of annihilating mankind [17]. For a discussion on these issues, we refer the reader to [56], [75], and [17].

Who or What Is Responsible When Robots Harm?

Veruggio [100], [102] dates the beginning of "roboethics" from two events. One was the Fukuoka World Robot Declaration, wherein it was stated that "next generation robots will contribute to the realization of a safe and peaceful society." The other was the roboethics road map [101], which sought to promote a cross-cultural discussion among scientists to monitor the effects of robotics technologies currently in use. More recently, an initial sketch of the code of ethics for the robotic community has been proposed [43]. This code offers general guidelines for ethical behavior. For example, the code reminds engineers that they may be held responsible for the actions of artificial creatures that they have helped to design. Along similar lines, Murphy and Woods [70] propose to rephrase the famous Asimov's laws, which they view as robot centric, in such a way as to remind robotics researchers and developers of their professional responsibilities. For example, the first law was replaced with "A human may not deploy a robot without the human—robot work system meeting the highest legal and professional standards of safety and ethics" [73, p. 19].

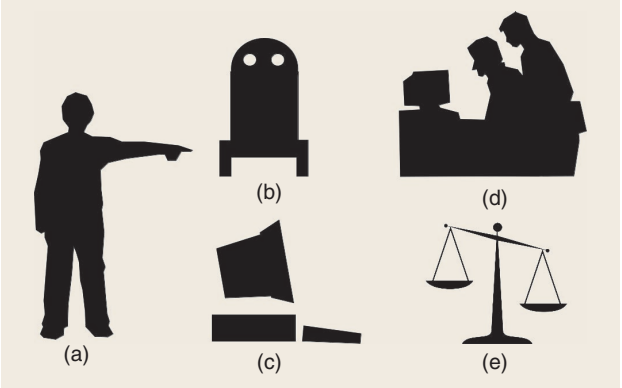

Figure 2. (a) The responsibility-ascription problem, i.e., the problem of assigning responsibility to the manufacturer, designer, owner, or user of the machine when use of this machine led to an armful event is a yet largely open issue. (b) People tend to blame the robots because they falsely attribute them with moral agency [29]. (c) People blame the machine even if they recognize the machine's lack of free will and lack of intentionality [28]. (d) Many ethicists argue that we should to some extent hold the engineers (the creators of the malfunctioning robots) responsible [60]. (e) To do so, we should use existing the legal principles, or create new ones, if necessary [13].

All the above implicates the responsibility ascription problem [69]: the problem of assigning responsibility to the manufacturer, designer, owner, or user of the robot or to the robot itself when using a robot leads to a harmful event. From a philosophical perspective, it is generally agreed that robots cannot themselves be held morally responsible [9], [25], [38] (although a few oppose this [95]) because computers as we conceive them today do not have intentionality [28]. From a psychological perspective, however, it remains an open question whether people include robots as an additional agent in the ascription of moral responsibility.

Who or what is responsible when robots harm (Figure 2)? Matthias [62] provides a seemingly simple answer. He argues that, in most cases, no one can be held accountable for the robotic failures. Matthias argues that with the advance of programming techniques (e.g., neural networks, evolutionary computation) that equip the agent with the ability to learn and, hence, to depart from its original

program, it becomes impossible for the programmer to exhaustively test the behaviors of his/her creations. In other words, the programmer can no longer foresee all possible sets of actions that the robot may take when in function. Hence, the programmer cannot be held responsible if harm should be done as a secondary effect of the robot interacting with humans, as long as the robot was not explicitly programmed to harm people. Matthias suggests that we should broadly adopt the idea of contracting insurances against harm caused by robots. Such a new type of insurance would ensure that, when no one can be held solely responsible for the harm done, then all the people involved in the incident would share the costs.

Marino and Tamburini [60] believe that Matthias's claims go too far. In their opinion, determining who is controlling the robot cannot be a criterion (albeit even the unique criterion) to ascribe responsibility. They argue that engineers cannot be freed from all responsibility on the sole ground that they do not have a complete control over the causal chains implied by the actions of their robots [60]. They rather offer to use legal principles that are routinely applied for other purposes, so as to fill the responsibility gap that Matthias emphasized. They take the example of the legislation in place for ascribing responsibility to the legally responsible person when harm is done by the dependent person. As a result, parents can be held responsible for the act of their children, when they can be found to have not provided adequate care or surveillance, even though there is no clear causal chain connecting them to the damaging events [63, p. 49]. A similar solution is proposed by Asaro [13], who draws a parallel between robots and any other completely unremarkable technological artifact[s] (e.g., a toaster or car). He shows that the Anglo-American civil law that rules for damages caused by these artifacts could also apply to damages produced by robots. For instance, if a manufacturer was aware of the danger that robots create, but failed to notify consumers, he may be charged with a failure to warn. And even if the producer did not know about the danger, he could be accused of failure to take proper care, meaning that the manufacturer failed to recognize some easily foreseeable threat brought upon by his/her technology.

On the downside, Asaro points out that, while the civil law can relatively be easily extended to rule for robot use, the criminal law is hardly applicable to the case of criminal actions caused by robots, as criminal actions can only be performed by moral agents. A moral agent is deemed so when it is recognized capable of understanding the moral concepts conveyed by the bylaws ruling our societies. Without a moral agency, the act of wrongdoing is considered an accident and not a crime. Furthermore, only a moral agent can be punished and reformed. This assumes that the moral agent has the ability to develop and correct its concept of morality [13]. In this context, the responsibility-ascription problem is, hence, reduced to the issue of attributing moral agency to the robot. Several authors have approached the problem of ascribing moral agency to robots [91]. For instance, Harnard [37] proposes to use some sort of moral Turing tests to establish whether the robot can be held responsible in court.

Another issue around the responsibility ascription problem centers on attributing moral agency to a robot. In one study, Friedman and Millett [30] found that 83% of the undergraduate computer science majors they interviewed attributed aspects of agency, either decision making or intentions, to computers. In addition, 21% of these students consistently held computers morally responsible for errors. In another article, Friedman and Kahn [28] identified a situation that may increase peoples attribution of agency to a machine, namely, when the machine is an expert recommendation system. Friedman and Kahn provide an example of the acute physiology and chronic health evaluation (APACHE) system [21]: a sophisticated computer-based modeling recommendation system to help hospital staff determine when to end life support for patients in intensive care units. Friedman and Kahn argue that the more such a system is relied on for objective and authoritative information, the more difficult it becomes to override its recommendations, and the more likely staff, including physicians, could begin to attribute moral agency toward the system. As a potential solution to such problems, Friedman and Kahn offer two design strategies. First, computational systems should be designed in ways that do not denigrate the human user to machinelike status. Second, computational systems should be designed in ways that do not impersonate human agency by attempting to mimic intentional states. The problem, however, in applying this second recommendation to robot design and implementation, especially those robots that have a humanoid form, is that such robots by design are conveying human attributes, thus fostering this problem.

Short- and long-term consequences of ethical issues are core to most of the current debates.

Ethical Issues in Service Robots

The design principle mentioned in the previous section aims at ensuring that robotic systems remain easily distinguishable from humans. Accordingly, this principle should help people ascribe responsibility in cases when the machine malfunctions or harms someone. However, as we noted, the current trend in robotics is the opposite, as there is a growing effort to design robots so that they look like humans [44], [45] or animals [31], [89].

The idea of designing machine-masquerading humans was questioned by Miller on the ground of human freedom [67]. Miller argues that, if humanlike robots really came to share the human space on a daily basis, the humans should be allowed to decide whether they wished to interact with these creatures; if they should decide they wanted to

interact solely with the other humans, they should be given the freedom to do so. Similarly, efforts at endowing robots with social skills have been criticized on the ground that the number of meaningful social interactions that humans that are typically capable to maintain is relatively small [23], [47]. Therefore, interacting with social artificial agents on a regular basis may lead people to become less prone to engage in social interactions with other people [66]. Others even hypothesized that people may come to build strong and perhaps even intimate bounds with robots and that this, again, may have negative side effects on the emotional relationships that people may be able to build with other people [50].

To shed some light on the aforementioned debate, people have started studying the type of human–robot relationships that arise when people interact with robotic systems that mimic human or animal behavior. In a series of four studies, Kahn and his colleagues studied children's social and moral relationships with the robot dog, the artificial intelligence robot (AIBO). The first three studies compared children's interaction with and reasoning about AIBO to, respectively, a stuffed (nonrobotic) dog [49], a biologically live dog [65], and a mechanical nonrobot dog [94], whereas the fourth study analyzed over postings in AIBO online discussion forums that spoke of members' relationships with their AIBO [30]. Together, these four studies provide converging evidence that children and adults can and often do establish meaningful and robust social conceptualizations and relationships with a robot that they recognize as a technology. For example, in the online discussion forum study, members affirmed that AIBO was a technology (75%), lifelike (48%), had mental states (60%), and was a social being (59%).

Robotic pets used in therapy with elderly may offer some level of companionship for which the elderly may be craving.

Across these four studies, however, the researchers found inconsistent findings in terms of people's commitments to AIBO as a moral agent. In an online discussion forum study, e.g., only 12% of the postings affirmed that AIBO had moral standing, including that AIBO had rights, merited respect, engendered moral regard, could be a recipient of care, or could be held morally responsible or blameworthy [30]. In contrast, in the Melson et al.'s [65] study, it was found that while, on the one hand, the children granted greater moral standing to a biologically live dog (86%) than to AIBO (76%), it was still striking that such a large percentage of children (76%) granted moral standing to the robot dog at all. One explanation for these inconsistent findings between studies is that the measures for establishing moral standing have been few and themselves difficult to interpret. For example, two of the five moral questions in the Melson et al.'s study were as follows: If you decided you did not like AIBO anymore is it OK or not OK to throw AIBO in the garbage? and If you decided you did not like AIBO anymore is it OK or not OK to destroy AIBO? The "not OK" answers were interpreted as indicating moral standing. However, one could plausibly make the same judgment about throwing away or destroying an expensive computer (because, e.g., it would wasteful) without committing morally to the artifact [65].

Since humans can develop emotional attachment toward robots, concerns have been expressed regarding the long-term consequences that such attachment may have on the individual. This is especially relevant when the person is fragile, as it is the case with children and people with mental delays. However, there are also several reasons to rather believe that interacting with social robots may benefit some of these individuals [48], [54], [97]. For instance, interacting with robots that display social behavior may help children with autism-impaired social skills [80], [26]. Robins et al. [80] conducted longitudinal studies over the course of several weeks of children with autism interacting with a humanoid robot. Unknown to the children, the robot was puppeteered so that it imitated the children's movement. Robins et al. showed that repeated exposure to the robot facilitated the emergence of spontaneous, proactive, and playful behavior, which these children very rarely display. Furthermore, once accustomed to the robot, the children also seem to engage in a more proactive interactive behavior with the adult investigator present in the room during the experiment. This leads, in some cases, to a triadic interaction: child–robot–adult. For example, children would acknowledge the presence of the investigator by spontaneously sitting on his/her lap for a few moments, holding his/her hand, or even trying to communicate by using simple words. However, it was not clear whether the social skills that children exhibited during the interactions with the robot had lasting effects.

In another study, Feil-Seifer and Mataric used a bubble-blowing robot in a three-some interaction child–caretaker–robot. While the robot was not actually behaving socially, its automatic bubble-blowing behavior provoked more child–caretaker interactions. In a similar triadic child–parent–robot scenario, Kozima and colleagues conducted a series of studies using Keepon, a simple two-link robot ball face, whose motions conveyed emotional expressions. These studies comfort Robins et al.'s findings that children with autism, in such a triadic scenario, spontaneously engage in social and affect display, which they otherwise tend to avoid [55], [26]. A comparative study of children with autism interacting with AIBO as opposed to a simpler mechanical toy showed enhanced verbal address directed to AIBO [94]. A survey of these studies can be found in [79].

As a whole, these studies seem to indicate that playing with robots that appear to behave in an autonomous and social manner may help children with autism-impaired more of these social skills that the autism therapy seeks to promote. Such a robotic-aided therapy does not aim

at developing attachment of the children toward the robot, but it might be a potential side effect. The question remains whether it is ethically correct to encourage children with autism to engage in affective interactions with machines incapable of emotions. Dautenhahn and Werry's response is that, "from the perspective of a person with autism and his/her needs, are these ethical concerns really relevant?"

Similarly, robotic pets used in therapy with elderly may offer some level of companionship. The seal robot, Paro, is probably the best example of such an application [89] [Figure 1(b)]. Wada et al. [104] reported on an extended use of Paro as part of therapeutic sessions in pediatric wards and elderly institutions worldwide. The results showed that the interaction with Paro improved the patients' and elderly people's moods and reduced their stress level [103]. It made them more active and communicative both among themselves and with their caretakers. A pilot study using electroencephalography (EEG) suggested that this robot therapy may improve the pattern of brain activity in patients suffering from dementia [104]. Furthermore, the effects of long-term interaction between Paro and the elderly were found to last for more than a year [105].

Although the aforementioned results speak in favor of using robots for therapy with the elderly, Sharkey offers a more cautious argumentation [85]. In his opinion, such surrogate companions do not really alleviate the elderly's isolation, and people are deluded about the real nature of their relationship to the devices [92] (Figure 3). Furthermore, even the robots that are clearly helping the elderly to maintain independence in their own homes [27] (e.g., robots used to remind the patient to take his/her medication) could lead to a situation where the elderly is left exclusively to the care of machines. However, the elderly's mental health substantially depends on human contact, which is to a large extent provided by the caregivers [93].

Figure 3. Interacting with robots that display social behavior may help children with autism-acquired social skills. The question remains whether it is ethically correct to encourage children with autism to engage in affective interactions with machines incapable of emotions. However, from the perspective of a person with autism, and his/her needs, are these ethical concerns really relevant? [23, p. 35]. In a broader context, some believe that the surrogate companions (e.g., robots assisting the elderly) are becoming more common because people are deluded about the real nature of their relationship to the devices [91]. (Photo courtesy of KASPAR robot by University of Hertfordshire [107].)

Robot nannies are another example of robotic applications that raise ethical questions [88]. There is an effort, mainly in South Korea and Japan, to build more sophisticated robots that could not only monitor babies [e.g., personal partner robot by National Electronics Conference (NEC) [32], Figure 1(a)] but would also be equipped with enough autonomy so as to call upon human caretakers only in unusual circumstances. It is likely that children will spend time playing with child-care robots, as researchers progress in designing ways for the robot to offer a sustained and rich interaction with the child, which may span months or even years [51], [63], [88]. This may, however, be detrimental to the physical and mental development of the child if children were to be left without human contact for many hours per day, as currently robotic pets are not designed to participate in the child's development in the same way as a child minder is trained to look after children [85]. This remains very speculative as the psychological impact that such robotics care may have on children's development is unknown. Some attempted to draw parallels with reports on severe social dysfunctions in young monkeys those interacted solely with artificial caretakers throughout the first years of development [61], [16], [88]. Perhaps of more pressing concern is the fact that there is no regulation to specifically deal with the case of child abuse when the child is cared for by a robot (national and international laws protecting children from mistreatment such as the United Nations Convention on the Rights of Child [71] do not cover this case) [88]. While one may argue that, when the time will really come to see robots caring for children, one will work on the associated legal issues, some people counter that this may be a bigger challenge than expected, as providing a unified code of ethics for regulating the use of robot nannies may be impossible owing to cultural differences between nations [36].

Ethical Issues in Lethal Robots

In the previous section, we discussed some of the ethical issues that stem from the current or foreseen robotic applications of service robots for education and therapy. Of equal if not more immediate ethical concerns are the current military applications of robots. Even though fully autonomous robots are not yet running in battlefields, as we will discuss here, the risks and benefits that introducing such autonomous lethal machine may have on wars are of crucial importance. Furthermore, because military technology often finds its way into civil applications, such as security or policing [14], [87], discussing the ethical issues related to military robots might also serve a broader context.

Currently, the decision to use a robotic device to kill human beings is still taken by a human operator. This decision stems from the desire to make sure that the human remains "in the loop," but it is not made out of technical

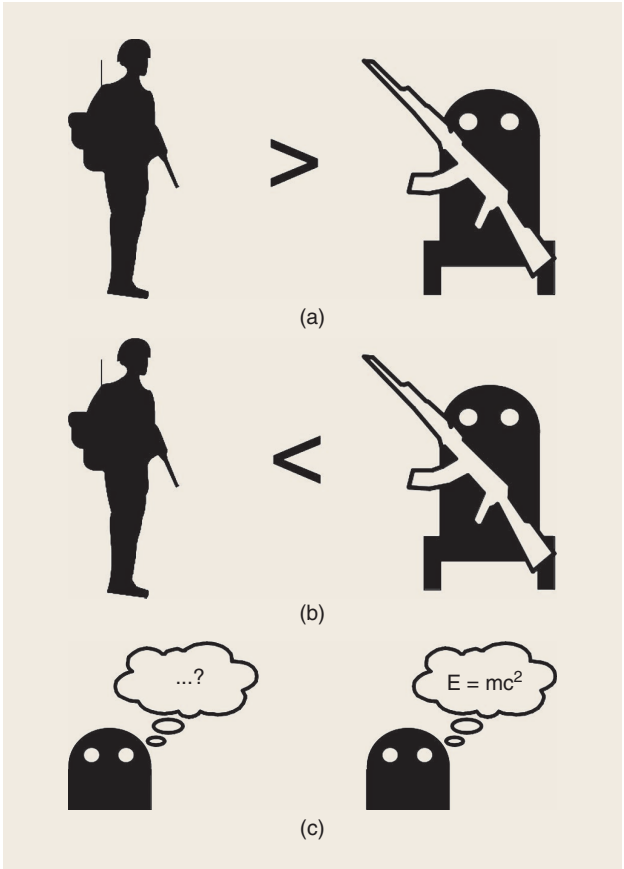

Figure 4. (a) Sharkey argues that the cognitive capabilities of robots do not match with that of humans, and thus lethal robots are unethical, as they may make mistakes more easily than humans [85]. (b) Arkin believes that although an unmanned system will not be able to perfectly behave in battlefield, it can perform more ethically than humans [9]. (c) In part, the question about the morality of using robots in the battlefield involves commitments on the capability of artificial intelligence. (Photo courtesy of the soldier's silhouette by Ruminglass and Quibik.)

necessity [14]. It is clear that the margin that separates us from having fully autonomous-armed systems in the battlefield is thinning. Even if all armed robots were to be supervised by humans, one may still wonder to what extent the human is still in control [9]. Moreover, there may be cases where one cannot avoid giving full autonomy to the system. For instance, combat aircrafts must be fully autonomous to effectively operate [99]. Sharkey predicts that, as the number of robots in operation in the battlefield increases, they may outnumber human soldiers. He then argues that it will become impossible for humans to simultaneously operate all these robots. Robots will then have to be fully autonomous [83].

One ethical issue (perhaps the issue that received most attention to date) arising from increasing autonomy of war robots has to do with the problem of discriminating between the fighters and innocent people. This distinction is at the core of the just war theory [106] and humanitarian laws [82]. These laws stipulate that only the fighters are legitimate targets and prohibit attacks against any other nonlegitimate targets [84], [14]. Sharkey rightfully argues that our robots are yet far from having visual capabilities that may allow to faithfully discriminate between the legitimate and nonlegitimate targets, even in close-contact encounter [85]. Besides, distinguishing between the legitimate and illegitimate targets is not purely technical and is further complicated by the lack of a clear definition of what is a civilian. (The 1944 Geneva Convention advises to use common sense, and the 1977 Protocol 1 defines a civilian as any person who is not a fighter [72].) However, even if one was provided with a precise definition that could be encoded in a computer program, it is doubtful that robots would achieve, in a foreseeable future, a level of complexity in robot cognition that would allow the robot to recognize ambiguous situations involving a nonlegitimate target manipulating lethal instruments (such as a situation where a child is carrying guns or ammunition). Sharkey argues that autonomous lethal systems should not be used, as long as one cannot fully demonstrate that the systems can faithfully distinguish between a soldier and civilian, and this in all situations [83]. Lin et al. believe that this is too stringent a condition, since even humans make errors of this kind (Figure 4) [58]. Arkin counters that, although unmanned robotic systems may make mistakes, it would on an average behave more ethically than human beings [9]. In support of this, Arkin cites the report from the Surgeon General's Office [96] regarding the ethics of soldiers. Less than half of the soldiers believed that the nonfighters should be treated with dignity. The other half was unclear as to how they should be treated. Moreover, one tenth of interrogated soldiers had mistreated nonfighters and one third reported having at least once faced a situation where they felt incapable of deciding what was the correct action (although all soldiers had received ethical training). Since human soldiers appear to misbehave from time to time, using machines that are more reliable and hence would, on average, make less mistakes should bring more good than harm. Lin et al. share the view that human soldiers are indeed less reliable and report on an evidence that human soldiers may act irrationally when in fear or stress. Hence, they concur that combat robots, which are affected by neither fear nor stress, may act more ethically than human soldiers irrespective of the circumstances [58].

Lin and colleagues point to one more issue related to using combat robots. As in the case of any other new technology, errors and bugs will inevitably exist, and these will lead combat robots to cause harmful accidents [58]. Such bugs or errors will be far more costly as human lives might be at stake. They advise to perform extensive testing of each military robot before usage. Nevertheless, they anticipate that, regardless of such efforts, combat robots may still occasionally behave in unexpected or unintended ways when used in the actual field [58]. Such errors could even lead to accidental wars if the robot's unexpected aggressive behavior was to be interpreted by the opponent as an act of war [14]. Groups of people interested in starting a war may seize upon such accidents to justify hostilities.

Even if one is not disputing the ethical question of fighting a war, one may want to question the ethics of having armed robots fully autonomous and used routinely in battlefields, especially when only one side may have robots. Politicians may tend to favor efforts made to replacing human fighters with robots, as each country feels a moral obligation to protect the lives of its soldiers [83]. However, there may be long-term consequences of waging these so-called risk-free wars ("A war where pilotless aircraft can beat a country's forces before sending in the ground robots to clean up" [87, p. 16]) or push-button wars ("A war in which the enemy is killed at a distance, without any immediate risk to oneself" [15, p. 62]). Since such wars will return wrecked metal instead of dead bodies (at least to the country using only robots), the emotional impact that wars currently have on civilians of that country will be largely lessened. The above is true only for the civilians not affected directly by combat, i.e., for wars fought in a distance.

It is feared that this may make it easier for a country to launch a war. These wars may also last for longer periods of time [58]. There are contradicting opinions whether this may result in people growing indifferent to the conduct of war. Sharkey fears that this would be the case [83], whereas Asaro believes that people are nearly always averse to starting an unjust war, irrespective of whether it would lead to human fatalities [15, p. 58]. That the war is risk free does not make it more acceptable [14]. Lin et al. counterweight this line of reasoning, arguing that such reasoning may lead to even more dangerously foolish ideas, such as the idea of trying to prevent wars to happen by increasing the brutality of fighting [58].

It was also argued that risk-free wars might increase terrorism, as the only possibility to strike back on a country that uses mainly robots in wars is to attack its citizens [83]. The less advanced, technologically speaking, side may advocate terrorism as a morally acceptable means to counterattack on the ground that robot armies are the product of a rich and elaborate economy, and that the members of that economy are the next-best legitimate targets [15, p. 64]. Hence, risk-free wars may paradoxically increase the risks for civilians [46]. However, Asaro reminds us that the wars are deemed morally acceptable as long as they do not harm civilians. According to this definition, terrorism would not be justified, irrespective of whether it is meant as a response to a country using robot armies. Thus, the fear that terrorism may increase as a result of using robot armies does not constitute, in Asaro's view, a valid moral objection to using robot armies. Only the questions of whether the robot armies can cause more harm or whether the use of such armies may lead to unjustified wars are of essence in the debate [14].

In contrast, Arkin anticipates that we will not end up with armies of unmanned systems operating on their own, but that rather heterogeneous teams composed of autonomous systems and humans soldiers will work together on the battlefield. He expects this to become a standard. Wars would, hence, not be fully risk free and so the dreaded consequences in increased terrorism or in societal indifference are not to be feared. Furthermore, Arkin expects that mixed teams, composed of robots and human soldiers, will act more ethically than groups composed of solely human soldiers. Robots equipped with video cameras (or other sensors) will record and report actions on the battlefield. Thus, they might serve as a deterrent against unethical behavior, as such acts would be registered. However, Lin and colleagues argue that if soldiers were to know that they are being watched by their fellow robot soldiers, they may no longer trust them and this could impact team cohesion. Consequently, human soldiers may fail to act adequately, e.g., by not providing support even if it is justified, out of stress caused by constant monitoring [58].

Robot nannies are another example of robotics applications that raise ethical questions.

Lastly, Sharkey points out that the legal status of war robots is unclear [86]. For example, while the unmanned aerial vehicle RQ-1 Predator [Figure 1(d)] was developed as a reconnaissance machine (hence the R in the name), it was subsequently equipped with hellfire missiles and renamed MQ-1 (where M stands for multipurpose). The MQ-1 was, however, never approved as a weapon. The fact of utmost concern is that, under current military standards, the MQ-1 does not need to be approved. Since the bare RQ-1 was not considered as a weapon (since it was meant only for surveillance) and that hellfire missiles have already been approved separately as weapons, the combination does not need special approval [19]. This may create a precedent whereby armed robots with growing level of autonomy can be created and used without any real legal control. In relation to legal issues, Asaro notes that "what is and what is not acceptable in war" is ultimately the subject of convention between nations [15, p. 64]. He argues that we can find support in existing laws only to certain extent. Eventually, the international community will be forced to create new laws and treaties to regulate the use of autonomous fighting robots.

Machine Ethics

Although still in its early stages, machine ethics offers a practical approach to introducing ethics in the design of autonomous machines. Machine ethics aims at giving the machine some autonomy, while ensuring that its behavior will abide ethical rules. Primarily, machine ethics seeks methods not only to ensure that the machine's behavior toward humans is proper [4], but it may also extend to designing rules driving ethical behavior of a machine toward another machine [6]. Machine ethics extends the field of computer ethics that is concerned with how people behave

with their computers to address the problem of how machines behave in general [2].

The interest in machine ethics is driven by the fact that robots have been already tightly integrated into human societies. Thus, since the robots already interact with humans and, as argued in the section "Who or What Is Responsible When Robots Harm?" engineers could be held responsible (to certain extent) for the actions of their creations; it is desirable to find methods of equipping the machines with moral behavior. Importantly, although the public attention might be focusing on the military application (such as Arkin's military adviser providing guidance on the use of lethal force by a robot [11]), machine ethics seems to be more concerned with service robots. There are many examples of such applications. Robots that share the workbench with humans in the industry might no longer be considered just a manufacturing tool but also as a "colleague" with whom workers interact [20]. Artificial sales agents in e-commerce, which can predict customers behaviors, should not abuse this knowledge by displaying unethical behavior [39]. Driverless trains in extreme situations might be forced to make decisions that could have life or death implications [2].

Although still in its early stages, machine ethics offers a practical approach to introducing ethics in the design of autonomous machines.

Asimov's laws of robotics are one of the first and best-known proposal to embed ethical concepts in the controller of the robot. (Asimov's laws of robotics were first introduced in the short science-fiction story Runaround [15].) According to these, all robots should under all circumstances obey three laws:

1) A robot may not injure a human being or, through inaction, allow a human being to be harmed.
2) A robot must obey orders it receives from human beings, except when such orders conflict with the first law.
3) A robot must protect its own existence as long as such protection does not conflict with the first or second law.

Later, Asimov added the fourth law (known as the law zero).

4) No robot may harm humanity or, through inaction, allow humanity to come to harm.

Many researchers recognize that Asimov's laws assume that robots have sufficient cognition to make moral decisions in all situations, including the complicated ones, in which even humans might have doubts [70]. Consequently, keeping in mind the current level of AI, these laws, although simple and elegant, serve no useful practical purpose [9] and are thus viewed as an unsatisfactory basis for machine ethics [8], [34]. Nevertheless, Asimov's laws often serve as a reference or starting point in the discussions related to machine ethics.

Fedaghi [1] proposes a classification scheme into ethical categories to simplify the process by which a robot may determine which action is most ethical in delicate situations. As a proof of concept, Fedaghi applies this classification to decompose Asimov's laws, hereby showing that these laws, once rephrased, can support logical reasoning. Such an approach is in line with the so-called procedural ethics [59], which develops procedures to guide the process by which ethical decisions are made [1]. A similar approach is presented in [18] that draws inspiration in Gottfried Wilhelm Leibniz's dream of a universal moral calculus [60]. There, deontic logic [22], [68] (i.e., logic extended with special operators for representing ethical concepts) is used instead of Asimov's laws to ground the robot's ethical reasoning. Such a methodology aims at maximizing the likelihood that a robot will behave in a certifiably ethical manner. That is, the robot's actions will be determined so that the ethical correctness of the resulting robot's behavior can be ensured through formal proofs. Such formal proofs check if a given robot 1) only takes permissible actions and 2) performs all obligatory actions (subject to ties and conflicts) [12]. Promoters of such methodology reason that human relationships and by extension human–robot relationships need to be based on some level of trust [107]. Such a formal and logical approach to describing robot behavior may help in determining whether the system is trustworthy. In contrast, they view inductive reasoning, which is based on case studies, as unreliable, because, while the "premise (success on trials) may all be true, the conclusion (desired behavior in the future) might still be false" [18], [90].

Others oppose this point of view and advocate the use of case-based reasoning (CBR) [74]. They reason that people can behave ethically without learning ethics (drawing a parallel to the fact that one can speak fluently a language without having received any formal grammar lessons) [81]. For example, McLaren implemented a CBR-ethical reasoner [64] and Anderson created a machine-learning system that automatically derives rules (principles) from cases provided by an expert ethicist [3], [7], [5]. For example, Arkin uses deliberative/reactive autonomous robotic architectures and provides the theory and formalisms for ethical control [10] and applies these to automatic military advisor [11]. He considers stimuli to behavior mappings and extends them with ethical constraints to ensure appropriate robot response (consistent with the law). In another example, Honarvar [40] used a CBR-like mechanism to train an artificial neural network to classify what is morally acceptable in a belief–desire–intention framework [77]. For example, he used this framework to augment the ethical knowledge of sales agent in an e-commerce application [39].

A particular machine ethics system that is very easy to implement is the one based on utilitarianism. It uses mathematical calculus to determine the best choice (by computing

and maximizing the goodness, however defined, of all actions) [4]. However, since utilitarianism values benefits brought upon society as a whole, hence ignoring the fate reserved to each individual in the society [78], such moral arithmetic cannot protect the fundamental rights of each individual [11] and as such is mostly of limited interest [35]. Still, practical work with a certain utilitarian flavor can be found in the literature, as most CBR systems previously presented assume that an arithmetic value is the main basis for determining what it is moral to do [53].

The last approach that we will mention is the rule-based one proposed by Powers. Powers argues that ethical systems such as Kant's categorical imperative naturally lead to a set of rules. (A categorical imperative denotes an absolute, unconditional requirement that asserts its authority in all circumstances, e.g., "act only according to that maxim whereby you can at the same time will that it should become a universal law" [55, p. 30].) This approach, hence, assumes that an ideological ethical code can be translated into a set of core rules. This is slightly similar to the deontic logic we reviewed earlier. It allows the robots to logically derive new ethical rules, appropriate to particular and new situations. Although interesting, this approach has not gathered much attention, as researchers usually turn to pure logic systems or CBR. In addition, Powers' ethical system had been criticized by Tonkens [98] on the basis that the development of Kantian artificial agents is itself against Kant's ethics. According to Kant, moral agents are both rational and free, whereas machines can only be rational. Hence, the mere fact of implementing a sense of morality into machines limits the machine's freedom of thought and reasoning.

In conclusion, machine ethics is composed of a number of interesting attempts to embed ethical rules in the robot's controller. These may be either popular ethics rules, such as Asimov's laws, or derived from classical philosophical approaches to ethics, such as Kant's ethics. Logical reasoning is the driving framework for most approaches. While still in infancy, machine ethics is a valuable attempt to conciliate the need to provide robots with ethical behavior with the need to make these machines more autonomous, as they come to support humans in their daily life. However, the approach may fall prey to several problems discussed throughout this article. Three of those stand out. One, if machines are not capable of being moral agents, as most philosophers agree, then it is important to design them with the ability to make moral decisions. Second, equipping the machines with morality (assuming it is possible) does not need to be a moral act on its own and might depend on the application one has in mind while developing a moral robot. For example, embedding morality into robot nannies or combat robots could lead to their widespread use, which could have severe negative consequences on the society. Finally, in an attempt to embed ethics into machines, because of their limited cognition, one must often unduly simplify the moral life. This seems to stand against the very goal of machine ethics itself (at least to some extent). It seems that it is still too early to judge whether the methods of machine ethics will prove useful or not and await more applications implemented in life.

Conclusions

Almost everyone agrees that they want robots to contribute to a better and more ethical world. The disagreements arise in how to bring that about. Some people want to embed ethical rules in the robots controller and employ such robots in morally challenging contexts, such as on the battlefield. Others argue vehemently against this approach: that robots themselves are incapable of being moral agents and thus should not be designed to have moral decision-making abilities. Others want to leverage the social aspects of robotics in bringing about human good. Along these lines, researchers have explored how robots can help children with autism or assist the elderly physically, thereby provide the elderly with enough autonomy to allow them to live in their own residence. Other researchers have explored how robots can provide companionship for the elderly and general population. Still others have worried that no matter how sophisticated robots become in their form and function, their technological platform will always distinguish people from them and prevent depth and authenticity of relation from forming. These are all open questions. Some are philosophical in nature, as is the question of whether robots are moral agents or could be in the future. Some are psychological, as in the question of whether people attribute moral responsibility to robots that harm. Some require political answers and new legislation. Finally, some, if not many, of the questions require thoughtful and on-going responses by those who engineer and design the robots. The engineer is also responsible for the ethical consequences of his/her creation. This seems at odds with the way research is currently done in robotics. Rarely, does one question the long-term ethical consequences of the research reported upon in scientific publications. (We are not referring here to short-term ethical consequences of a research, such as a research that involves human subjects. Clearly, these are always carefully scrutinized, and this research must be approved by the ethical committee before the conduct of the project.) There are several reasons for this. On the one hand, most of these damaging long-term consequences seem very speculative and still far away from the technological reality. On the other hand, it is expected that these issues will be disputed at a political level, and, hence, that it is perhaps not the role of the engineers and scientists to discuss these.

Some scientists, however, discuss these issues, but, as with any debate, people sometimes have opposite views on which robotic application is ethical and which is not. We showed that such dissensions stemmed often from different beliefs on human nature and different expectations on what technology may achieve in the future. Although it is difficult to anticipate how and when robots will come to play an active role in our society, there is no reason why one should

not continue discussing various scenarios. We might be motivated by the beauty of our artifacts, their usefulness, or the economic rewards. However, in addition, we are morally accountable for what we design and put out into the world.

Acknowledgments

This work was partially supported by the Swiss National Science Foundation (grant number K-23K0-117914), the European Commission under contract number FP7-248258 (First-MM), and the National Science Foundation in the United States (grant number IIS-0905289).

References

[1] S. S. Al-Fedaghi, "Typification-based ethics for artificial agents," in *Proc. 2nd IEEE Int. Conf. Digital Ecosystems and Technologies (DEST)*, 2008, pp. 482–491.

[2] C. Allen, W. Wallach, and I. Smit, "Why machine ethics?" *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 12–17, 2006.

[3] M. Anderson, S. L. Anaderson, and C. Armen, "MedEthEx: A prototype medical ethics advisor," in *Proc. 18th Conf. Innovative Applications of Artificial Intelligence (IAAI)*, 2006, pp. 1759–1765.

[4] M. Anderson, S. Anderson, and C. Armen, "Towards machine ethics: Implementing two action-based ethical theories," in *Proc. AAAI Fall Symp. Machine Ethics*, 2005, pp. 1–7, Tech. Rep. FS-05-06.

[5] M. Anderson and S. L. Anderson, "Ethical healthcare agents," in *Advanced Computational Intelligence Paradigms in Healthcare—3* (Stud. Comput. Intell., vol. 107), M. Sordo, S. Vaidya, and L. C. Jain, Eds. Berlin: Springer-Verlag, 2008, pp. 233–257.

[6] M. Anderson and S. L. Anderson, "Machine ethics: Creating an ethical intelligent agent," *AI Mag.*, vol. 28, no. 4, pp. 15–27, 2007.

[7] M. Anderson, S. L. Anderson, and C. Armen, "An approach to computing ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 56–63, 2006.

[8] S. L. Anderson, "Asimov's 'three laws of robotics' and machine metaethics," *AI Soc.*, vol. 22, no. 4, pp. 477–493, 2008.

[9] R. C. Arkin, "Governing ethical behavior: Embedding an ethical controller in a hybrid deliberative-reactive robot architecture—Part I: Motivation and philosophy," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interaction*, 2008, pp. 121–128.

[10] R. C. Arkin, "Governing ethical behavior: Embedding an ethical controller in a hybrid deliberative-reactive robot architecture—Part II: Formalization for ethical control," in *Proc. 1st Conf. Artificial General Intelligence*, 2008, pp. 51–62.

[11] R. C. Arkin. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative-reactive robot architecture—Part III: Representational and architectural considerations. presented at Proc. Technology in Wartime Conf. Stanford Law School [Online]. Available: http://hdl.handle.net/1853/22715

[12] K. Arkoudas, S. Bringsjord, and P. Bello, "Toward ethical robots via mechanized deontic logic," in *Proc. AAAI Fall Symp. Machine Ethics*, 2005, Tech. Rep. FS-05-06, pp. 17–23.

[13] P. Asaro. (2007). Robots and responsibility from a legal perspective. presented at Proc. Workshop on Roboethics, IEEE Int. Conf. Robotics and Automation (ICRA) [Online]. Available: http://www.roboethics.org/icra2007/contributions/ASARO LegalPerspective.pdf

[14] P. Asaro, *How Just Could a Robot War Be?* Amsterdam: IOS Press, 2008, pp. 50–64.

[15] I. Asimov, "Runaround," *Astounding Science Fiction*. New York: Dell Magazines, 1942.

[16] D. Blum, *Love at Goon Park: Harry Harlow and the Science of Affection*. New York: Basic Books, 2002.

[17] N. Bostrom. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *J. Evol. Technol.*, [Online]. *9(1)*. Available: http://www.jetpress.org/volume9/risks.html

[18] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a general logicist methodology for engineering ethically correct robots," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 38–44, 2006.

[19] J. Canning, G. Riggs, O. Holland, and C. Blakelock, "A concept for the operation of armed autonomous systems on the battlefield," in *Proc. Association for Unmananned Vehicle Systems Int. Annu. Symp. and Exhibition*, Anaheim, CA.

[20] B. Curuklu, G. Dodig-Crnkovic, and B. Akan, "Towards industrial robots with human-like moral responsibilities," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2010, pp. 85–86.

[21] R. W. S. Chang, B. Lee, S. Jacobs, and B. Lee, "Accuracy of decisions to withdraw therapy in critically ill patients: Clinical judgment versus a computer model," *Crit. Care Med.*, vol. 17, no. 11, pp. 1091–1097, 1989.

[22] B. F. Chellas, *Modal Logic: An Introduction*. Cambridge, MA: Cambridge Univ. Press, 1980.

[23] K. Dautenhahn, "Robots we like to live with?!—A developmental perspective on a personalized, life-long robot companion" in *Proc. 13th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*, 2004, pp. 17–22.

[24] K. Dautenhahn and I. Werry, "Towards interactive robots in autism therapy: Background, motivation and challenges," *Pragmat. Cognition*, vol. 12, no. 1, pp. 1–35, 2004.

[25] D. Dennett, *When HAL Kills, Who's to Blame?* Cambridge, MA: MIT Press, 1996, ch. 16.

[26] D. Feil-Seifer and M. Mataric, "Robot-assisted therapy for children with autism spectrum disorders," in *Proc. 7th Int. Conf. Interaction Design and Children*, 2008, pp. 49–52.

[27] J. Forlizzi, C. DiSalvo, and F. Gemperle, "Assistive robotics and an ecology of elders living independently in their homes," *Hum. Comput. Interact.*, vol. 19, no. 1, pp. 25–59, 2004.

[28] B. Friedman, "'It's the computer's fault': Reasoning about computers as moral agents," in *Proc. Conf. Companion on Human Factors in Computing Systems*, 1995, pp. 226–227.

[29] B. Friedman and P. H. Kahn, Jr., "Human agency and responsible computing: Implications for computer system design," *J. Syst. Softw.*, vol. 17, no. 1, pp. 7–14, 1992.

[30] B. Friedman, P. H. Kahn, Jr., and J. Hagman, "Hardware companions?: What online AIBO discussion forums reveal about the human-robotic relationship" in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2003, pp. 273–290.

[31] M. Fujita, "AIBO: Toward the era of digital creatures," *Int. J. Robot. Res.*, vol. 20, no. 10, pp. 781–794, 2001.

[32] Y. Fujita, S. I. Onaka, Y. Takano, J. U. N. I. Funada, T. Iwasawa, T. Nishizawa, T. Sato, and J. U. N. I. Osada, (2005). "Development of child-care robot PaPeRo," *Nippon Robotto Gakkai Gakujutsu Koenkai Yokoshu*, [CD-ROM], *23*, pp. 1–11. Available: http://sciencelinks.jp/j-east/article/200523/000020052305A0951578.php

[33] General Atomics Aeronautical. (2010, Sept. 17). Predator B [Online]. Available: http://www.ga-asi.com/products/aircraft/predator_b.php

[34] J. Gips, *Towards the Ethical Robot*. Cambridge, MA: MIT Press, 1995, pp. 243–252.
[35] C. Grau, "There is no 'I' in 'robot': Robots and utilitarianism," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 52–55, 2006.
[36] S. Guo and G. Zhang, "Robot rights," *Science*, vol. 323, no. 5916, p. 876, 2009.
[37] S. Harnad, "Minds, machines and Turing," *J. Logic Lang. Inform.*, vol. 9, no. 4, pp. 425–445, 2000.
[38] K. E. Himma, "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics Inform. Technol.*, vol. 11, no. 1, pp. 19–29, 2009.
[39] A. R. Honarvar and N. Ghasem-Aghaee, "Towards an ethical sales-agent in e-commerce," in *Proc. 2010 Int. Conf. e-Education, e-Business, e-Management and e-Learning*, 2010, pp. 230–233.
[40] A. R. Honarvar and N. Ghasem-Aghaee, "An artificial neural network approach for creating an ethical artificial agent," in *Proc. 8th IEEE Int. Conf. Computational Intelligence in Robotics and Automation*, 2009, pp. 290–295.
[41] IFR Statistical Department. (2010, May 5). Executive summary of world robotics 2009 industrial robots and service robots [Online]. Available: http://www.worldrobotics.org/downloads/2009_executive_summary.pdf
[42] Foster-Miller Inc. (2010, Sept. 17). TALON Family of Military, Tactical, EOD, MAARS, CBRNE, Hazmat, SWAT and Dragon Runner Robots [Online]. Available: http://foster-miller.qinetiq-na.com/lemming.htm
[43] B. Ingram, D. Jones, A. Lewis, M. Richards, C. Rich, and L. Schachterle, "A code of ethics for robotics engineers," in *Proc. 5th ACM/IEEE Int. Conf. Human–Robot Interaction*, 2010, pp. 103–104.
[44] H. Ishiguro, "Android science: Conscious and subconscious recognition," *Connection Sci.*, vol. 18, no. 4, pp. 319–332, 2006.
[45] H. Ishiguro, "Interactive humanoids and androids as ideal interfaces for humans," in *Proc. 11th Int. Conf. Intelligent User Interfaces*, 2006, pp. 2–9.
[46] P. H. Kahn, Jr., "The paradox of riskless warfare," *Philos. Public Policy Quart.*, vol. 22, no. 3, pp. 2–8, 2002.
[47] P. H. Kahn, Jr., N. G. Freier, B. Friedman, R. L. Severson, and E. Feldman, "Social and moral relationships with robotic others," in *Proc. 13th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*, 2004, pp. 545–550.
[48] P. H. Kahn, Jr., B. Friedman, and J. Hagman, "I care about him as a pal: Conceptions of robotic pets in online AIBO discussion forums," in *Proc. CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 2002, pp. 632–633.
[49] P. H. Kahn, Jr., B. Friedman, D. R. Perez-Granados, and N. G. Freier, "Robotic pets in the lives of preschool children," in *Proc. CHI'04 Extended Abstracts on Human Factors in Computing Systems*, 2004, pp. 1449–1452.
[50] P. H. Kahn, Jr., J. H. Ruckert, T. Kanda, H. Ishiguro, A. Reichert, H. Gary, and S. Shen, "Psychological intimacy with robots?: Using interaction patterns to uncover depth of relation" in *Proc. 5th ACM/IEEE Int. Conf. Human–Robot Interaction*, 2010, pp. 123–124.
[51] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Hum. Comput. Interact.*, vol. 19, no. 1–2, pp. 61–84, 2004.
[52] I. Kant (Transl.: J. W. Ellington), *Grounding for the Metaphysics of Morals*, 3rd ed. Indianapolis: Hackett Pub. Co., 1993 (written in 1785).
[53] M. Keefer, *Moral Reasoning and Case-Based Approaches to Ethical Instruction in Science* (Science and Technology Education Library 19). New York: Springer-Verlag, 2003, pp. 241–259.
[54] C. Kidd, W. Taggart, and S. Turkle, "A sociable robot to encourage social interaction among the elderly," in *Proc. 2006 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2006, pp. 1050–4729.
[55] H. Kozima, C. Nakagawa, and Y. Yasuda, "Children–robot interaction: A pilot study in autism therapy," in *From Action to Cognition* (Prog. Brain Res., vol. 164), C. von Hofsten and K. Rosander, Eds., 2007, pp. 385–400.
[56] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking Penguin, 2005.
[57] G. W. Leibniz (Transl.: G. M. Ross), *Notes on Analysis* (Past Masters). London, U.K.: Oxford Univ. Press, 1984.
[58] P. Lin, G. A. Bekey, and K. Abney, "Robots in war: Issues of risk and ethics," in *Ethics and Robotics*. Amsterdam: IOS Press, 2009, pp. 49–67.
[59] W. Maner, "Heuristic methods for computer ethics," *Metaphilosophy*, vol. 33, no. 3, pp. 339–365, 2002.
[60] D. Marino and G. Tamburrini, "Learning robots and human responsibility," *Int. Rev. Inform. Ethics*, vol. 6, no. 006, pp. 46–50, 2006.
[61] W. A. Mason and G. Berkson, "Effects of maternal mobility on the development of rocking and other behaviors in rhesus monkeys: A study with artificial mothers," *Develop. Psychobiol.*, vol. 8, no. 3, pp. 197–211, 1975.
[62] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics Inform. Technol.*, vol. 6, no. 3, pp. 175–183, 2004.
[63] N. Mavridis, C. Datta, S. Emami, A. Tanoto, C. BenAbdelkader, and T. Rabie, "FaceBots: Robots utilizing and publishing social information in facebook," in *Proc. 4th ACM/IEEE Int. Conf. Human Robot Interaction*, 2009, pp. 273–274.
[64] B. M. McLaren, "Computational models of ethical reasoning: Challenges, initial steps, and future directions," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 29–37, 2006.
[65] G. F. Melson, P. H. Kahn, Jr., A. Beck, and B. Friedman, "Robotic pets in human lives: Implications for the human–animal bond and for human relationships with personified technologies," *J. Soc. Issues*, vol. 65, no. 3, pp. 545–567, 2009.
[66] G. F. Melson, P. H. Kahn, Jr., A. M. Beck, B. Friedman, T. Roberts, and E. Garrett, "Robots as dogs?: Children's interactions with the robotic dog AIBO and a live Australian shepherd" in *Proc. CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1649–1652.
[67] K. W. Miller, "It's not nice to fool humans," *IT Prof.*, vol. 12, no. 1, pp. 51–52, 2010.
[68] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, 2006.
[69] P. A. Mudry, S. Degallier, and A. Billard, "On the influence of symbols and myths in the responsibility ascription problem in roboethics—A roboticist's perspective," in *Proc. 17th IEEE Int. Symp. Human Robot Interactive Communication (RO-MAN)*, 2008, pp. 563–568.
[70] R. Murphy and D. D. Woods, "Beyond Asimov: The three laws of responsible robotics," *IEEE Intell. Syst.*, vol. 24, no. 4, pp. 14–20, 2009.
[71] United Nations. (1989). Convention on the rights of the child, *Treaty Series*, [Online]. 1577, p. 3. Available: http://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-11&chaptez=4&lang=en

[72] International Committee of the Red Cross. (1977). Protocol additional to the Geneva conventions of 12 August 1949 [Online]. Article 50. Available: http://www.icrc.org/ihl.nsf/7c4d08d9b287a42141256739003e636b/f6c8b9fee14a77fdc125641e0052b079

[73] J. Osada, S. Ohnaka, and M. Sato, "The scenario and design process of childcare robot, PaPeRo," in *Proc. 2006 ACM SIGCHI Int. Conf. Advances in Computer Entertainment Technology*, 2006, pp. 80–87.

[74] S. K. Pal and S. Shiu, *Foundations of Soft Case-Based Reasoning*. Hoboken, NJ: Wiley-Interscience, 2004.

[75] S. Petersen, "The ethics of robot servitude," *J. Exp. Theor. Artif. Intell.*, vol. 19, no. 1, pp. 43–54, 2007.

[76] D. Pressler. (2010, Sept. 23). Stupid little robot buddy [Online]. Available: http://www.davepresslerart.com/

[77] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in *Proc. 1st Int. Conf. Multiagent Systems (ICMAS)*, 1995, pp. 312–319.

[78] J. Rawls, *A Theory of Justice*. Cambridge: Harvard Univ. Press, 1999.

[79] D. J. Ricks and M. B. Colton, "Trends and considerations in robot-assisted autism therapy," in *Proc. 2010 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2010, pp. 4354–4359.

[80] B. Robins, K. Dautenhahn, R. T. Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?" *Univers. Access Inform. Soc.*, vol. 4, no. 2, pp. 105–120, 2005.

[81] R. Rzepka and K. Araki, "What could statistics do for ethics? The idea of a commonsense processing based safety valve" in *Proc. AAAI Fall Symp. Machine Ethics*, 2005, pp. 85–87, Technical Report FS-05-06.

[82] M. N. Schmitt, "The principle of discrimination in 21st century warfare," *Yale Hum. Rights Develop. Law J.*, vol. 2, no. 1, pp. 143–164, 1999.

[83] N. Sharkey, "Cassandra or false prophet of doom: AI robots and war," *IEEE Intell. Syst.*, vol. 23, no. 4, pp. 14–17, 2008.

[84] N. Sharkey, "Grounds for discrimination: Autonomous robot weapons," *RUSI Defence Syst.*, vol. 11, no. 2, pp. 86–89, 2008.

[85] N. Sharkey, "The ethical frontiers of robotics," *Science*, vol. 322, no. 5909, pp. 1800–1801, 2008.

[86] N. Sharkey, "Death strikes from the sky: The calculus of proportionality," *IEEE Technol. Soc. Mag.*, vol. 28, no. 1, pp. 16–19, 2009.

[87] N. Sharkey, "The robot arm of the law grows longer," *Computer*, vol. 42, no. 8, pp. 113–116, 2009.

[88] N. Sharkey and A. Sharkey, "The crying shame of robot nannies: An ethical appraisal," *Interact. Stud.*, vol. 11, no. 2, pp. 161–190, 2010.

[89] T. Shibata, M. Yoshida, and J. Yamato, "Artificial emotional creature for human-machine interaction," in *Proc. 1997 IEEE Int. Conf. Systems, Man, and Cybernetics*, 1997, vol. 3, pp. 2269–2274.

[90] B. Skyrms, *Choice and Chance: An Introduction to Inductive Logic*. Belmont, CA: Wadsworth, 1999.

[91] L. B. Solum, "Legal personhood for artificial intelligences," *North Carolina Law Rev.*, vol. 70, no. 1, pp. 1231–1287, 1992.

[92] R. Sparrow, "The march of the robot dogs," *Ethics Inform. Technol.*, vol. 4, no. 4, pp. 305–318, 2002.

[93] R. Sparrow and L. Sparrow, "In the hands of machines? The future of aged care" *Minds Mach.*, vol. 16, no. 2, pp. 141–161, 2006.

[94] C. M. Stanton, P. H. Kahn, Jr., R. L. Severson, J. H. Ruckert, and B. T. Gill, "Robotic animals might aid in the social development of children with autism," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interaction*, 2008, pp. 271–278.

[95] J. P. Sullins, "When is a robot a moral agent?" *Int. Rev. Inform. Ethics*, vol. 6, no. 12, pp. 23–30, 2006.

[96] Surgeon General's Office. (2006, Nov. 17). Mental health advisory team (MHAT) IV operation Iraqi freedom 05-07. Final Report [Online]. Available: http://www.armymedicine.army.mil/reports/mhat/mhat_iv/MHAT_IV_Report_17NOV06.pdf

[97] A. Tapus, C. Tapus, and M. J. Matarić, "Music therapist robot for individuals with cognitive impairments," in *Proc. 4th ACM/IEEE Int. Conf. Human Robot Interaction*, 2009, pp. 297–298.

[98] R. Tonkens, "A challenge for machine ethics," *Minds Mach.*, vol. 19, no. 3, pp. 421–438, 2009.

[99] US Department of Defense, Office of the Assistant Secretary of Defense (Public Affairs) News Transcript. (2007, Dec. 18). DoD Press Briefing with Mr. Weatherington from the Pentagon Briefing Room, Arlington, VA [Online]. Available: http://www.defense.gov/Transcripts/Transcript.aspx?TranscriptID=4108

[100] G. Veruggio. (2005). The birth of roboethics. presented at Workshop on Robo-Ethics, IEEE Int. Conf. Robotics and Automation (ICRA) [Online]. Available: http://www.roboethics.org/icra2005/veruggio.pdf

[101] G. Veruggio, "The EURON roboethics roadmap," in *Proc. 2006 6th IEEE-RAS Int. Conf. Humanoid Robots*, 2006, pp. 612–617.

[102] G. Veruggio and F. Operto, "Roboethics: A bottom-up interdisciplinary discourse in the field of applied ethics in robotics," *Int. Rev. Inform. Ethics*, vol. 6, no. 006, pp. 2–8, 2006.

[103] K. Wada and T. Shibata, "Living with seal robots its sociopsychological and physiological influences on the elderly at a care house," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 972–980, 2007.

[104] K. Wada, T. Shibata, T. Musha, and S. Kimura, "Robot therapy for elders affected by dementia," *IEEE Eng. Med. Biol.*, vol. 27, no. 4, pp. 53–60, 2008.

[105] K. Wada, T. Shibata, T. Saito, K. Sakamato, and K. Tanie, "Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged," in *Proc. 2005 IEEE Int. Conf. Robotics and Automation (ICRA)*, 2005, pp. 2785–2790.

[106] M. Walzer, *Just and Unjust Wars*, 3rd ed. New York: Basic Books, 2000.

[107] V. Wiegel and J. van den Berg, "Combining moral theory, modal logic and MAS to create well-behaving artificial agents," *Int. J. Soc. Robot.*, vol. 1, no. 3, pp. 233–242, 2009.

[108] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A.f Mirza, and M. Blow, "KASPAR—A minimally expressive humanoid robot for human-robot interaction research," *Appl. Bionics Biomech.*, vol. 6, no. 3, pp. 369–397, 2009.

Pawel Łichocki, Laboratory of Intelligent Systems, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland. E-mail: pawel.lichocki@epfl.ch.

Peter H. Kahn, Jr., Department of Psychology, University of Washington, Seattle, WA 98195-1525. E-mail: pkahn@u.washington.edu.

Aude Billard, Learning Algorithms and Systems Laboratory, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland. E-mail: aude.billard@epfl.ch.