Ethics Inf Technol (2012) 14:61–71
DOI 10.1007/s10676-011-9278-2

ORIGINAL PAPER

Robots: ethical by design

Gordana Dodig Crnkovic · Baran Çürüklü

Published online: 24 August 2011

Abstract Among ethicists and engineers within robotics there is an ongoing discussion as to whether ethical robots are possible or even desirable. We answer both of these questions in the positive, based on an extensive literature study of existing arguments. Our contribution consists in bringing together and reinterpreting pieces of information from a variety of sources. One of the conclusions drawn is that artifactual morality must come in degrees and depend on the level of agency, autonomy and intelligence of the machine. Moral concerns for agents such as intelligent search machines are relatively simple, while highly intelligent and autonomous artifacts with significant impact and complex modes of agency must be equipped with more advanced ethical capabilities. Systems like cognitive robots are being developed that are expected to become part of our everyday lives in future decades. Thus, it is necessary to ensure that their behaviour is adequate. In an analogy with artificial intelligence, which is the ability of a machine to perform activities that would require intelligence in humans, artificial morality is considered to be the ability of a machine to perform activities that would require morality in humans. The capacity for artificial (artifactual) morality, such as artifactual agency, artifactual responsibility, artificial intentions, artificial (synthetic) emotions, etc., come in varying degrees and depend on the type of agent. As an illustration, we address the assurance of safety in modern High Reliability Organizations through responsibility distribution. In the same way that the concept of *agency* is generalized in the case of *artificial agents*, the concept of *moral agency,* including *responsibility,* is generalized too. We propose to look at artificial moral agents as having functional responsibilities within a network of distributed responsibilities in a socio-technological system. This does not take away the responsibilities of the other stakeholders in the system, but facilitates an understanding and regulation of such networks. It should be pointed out that the process of development must assume an evolutionary form with a number of iterations because the emergent properties of artifacts must be tested in real world situations with agents of increasing intelligence and moral competence. We see this paper as a contribution to the macro-level Requirement Engineering through discussion and analysis of general requirements for design of ethical robots.

Keywords Artificial morality · Machine ethics · Machine morality · Roboethics · Autonomous agents · Artifactual responsibility · Functional responsibility

Introduction

Robots as intelligent agents are one of the most promising future emerging technologies (Gates 2007; Warwick 2009). The more intelligent they become the more useful and effective they are. However, historical experience shows that highly intelligent agents without ethical qualities may easily turn out to be unscrupulous and destructive. The purpose of this article is to show why and how ethics should enter the field of intelligent robots/softbots and contribute to the promotion of the idea that *intelligence*

G. Dodig Crnkovic (✉)
Computer Science Laboratory, School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden
e-mail: gordana.dodig-crnkovic@mdh.se

B. Çürüklü
Computational Perception Laboratory, School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden
e-mail: baran.curuklu@mdh.se

must come in conjunction with ethics, through the concept of an artifact ethical by design.

Autonomous AI agents′ ethical aspects have been insufficiently researched until now, among others based on the misconception that intelligent artifacts do essentially what they have been programmed to do (Lin et al. 2008) p. 8, which is true only for very simple agents. With growing complexity and increasing autonomy, learning and adaptive abilities; ethical challenges are multiplying. They include engineering ethics of designers, manufacturers, and maintenance services, as well as ethical attitudes of users and ethical aspects of the artifacts themselves.

At the moment there is a great conceptual confusion in understanding the implications of artifacts with cognitive capacities. They present cognitive extensions that increase our knowledge and support our agency. As we progress, we shape the technology, which, in turn, shapes us. Intelligent artifacts will change future human society. Consequently, it is worthwhile to understand the possible options in a time perspective of next few decades, when intelligent artifacts are expected to enter many social spheres—from entertainment to medicine and elderly care, to schools, industry and infrastructures. Those developments have ethical consequences that should be analyzed and understood proactively. We are gradually becoming prepared for this new stage with cognitive robots/softbots in the society. Only a collaborative effort across disciplines can provide relevant insights into the complexity of integration of a "new, intelligent, artificial species" among us.

An interesting aspect of the development of cognitive machines with "built-in" Machine Ethics is the prospect of a deeper understanding of the mechanisms of ethical behaviour in humans. In the same way that we learned about human intelligence by building artificial intelligent agents, we may expect to learn about ethics by building ethical agents (Allen et al. 2006). Historically, while building intelligent machines, important new facets of intelligence emerged—its embodiment and embeddedness. Newly, synthetic emotions are being developed for implementation in the machines, (Vallverdú and Casacuberta 2009). This adds a new dimension to the human–machine interaction.

Ethics of artifactually intelligent robotic agents

The ethics of robotic agents is the subject of two main computer ethics fields. First is Engineering Ethics, which, in the first place puts responsibility on the engineers involved, hoping that they will retain full control over the artifacts, no matter how complex or autonomous. Second is Machine Ethics, which argues that ethics should be designed into intelligent artifacts that are required to behave autonomously according to ethical standards.

Among the pioneers of the Machine Ethics field are Wallach and Allen (2009), Anderson and Anderson (2007), Floridi and Sanders (2004), Moor (2006), Magnani (2007), Scheutz (2002), Sullins (2006), and Edgar (1997). The most prominent representatives of Engineering Ethics within Computing include Moor (1985), Mitcham (1995), Bynum and Rogerson (2004), Johnson (1994), Johnson and Miller (2006).

Robotic Ethics approaches robots using both Machine Ethics and Engineering Ethics. Notable contributors include Arkin (1998), Veruggio and Operto (2008), Beavers (2011), Capurro and Nagenborg (2009), Clark (2003), Coeckelbergh (2009), Levy (2006), Lin et al. (2009), Moravec (1999), and Nagenborg (2007). Roboethics has brought into focus numerous ethical aspects of robotics, including designers, manufacturers, and users of robots (Veruggio 2006). It has a time perspective of a few decades in order to avoid speculation and excessive uncertainty. From what we know today, conscious robots do not belong to the near future and are not in the focus of roboethics. The emphasis is on the phenomena in the domain of applied ethics, relying on already existing insights that are in a sense close to the Computer Ethics of James Moor, Terrell Bynum and Deborah Johnson. Based on the existing technology, the dominant view is that:

> Robots are and will remain in the foreseeable future dependent on human ethical scrutiny as well as on the moral and legal responsibility of humans. (Capurro and Nagenborg 2009)

Obviously designers of intelligent, autonomous robots bear special responsibility for their functioning and appropriate behaviour. (Miller 2011) Among the core Requirement Engineering activities (Nuseibeh and Easterbrook 2000)—eliciting, modeling and analysing, communicating, agreeing, and evolving requirements—of special interest for us is the eliciting, discussing and analyzing requirements necessary to assure robots ethical by design. Requirement Engineering is both a macro-level organizational activity (in deciding what sort of requirements will go into products) and a micro-level project activity concerning the final requirements (Aurum and Wohlin 2003). There are authors who already discuss architectures for the implementation of machine morality (Allen et al. 2005; Anderson and Anderson 2007) and in their case one could discuss the whole spectrum of Requirements Engineering activities. However, we will remain on the macro-level requirements, arguing for the necessity of intelligent autonomous artifacts that have moral behaviour by design (i.e., by construction).

This is a debated issue because of the concern that building moral characteristics, such as building (artificial)

responsibility into artifacts, may result in humans not taking responsibility and blaming the artifacts for all possible problems. However, as Verbeek emphasizes:

> Engineering ethics mainly focuses on the moral decisions and responsibilities of designers, and remains too external to the moral significance of technologies themselves. Yet, analyses of the non-neutrality of technology make it plausible to ascribe some morality to artifacts. (Verbeek 2008)

Engineering ethics and Science and Technology Studies (STS) have by tradition developed independently, while (Van de Poel and Verbeek 2006), together with other contributors of the special issue of *Technology and Human Values* on *Ethics and Engineering Design,* argue that those two fields have a lot to learn from each other. This view is supported by Brey's account of technology as the outcome of evolutionary processes rather than of intelligent design (Brey 2008).

In other words, the design is one of the essential contributions that define an artifact, along with implementation (use). It is thus important to elucidate why intrinsic ethical behaviour built into robots cannot remove the responsibility from designers, manufacturers, and other stakeholders (Davis 2010). In the coming chapters, we will give an example of moral responsibility distribution in High Reliability Organizations.

Dodig-Crnkovic and Persson (2008) present an argument that socio-technological systems must be viewed as networks of distributed moral responsibility, where responsibility for a task with moral significance can be seen as moral responsibility even when an agent who is responsible for a task is an intelligent machine. We suggest that what should be compared from an ethical point of view is the behaviour of an intelligent autonomous agent without any ethical capacity and an equivalent one with artifactual morality. One may see morality as added value that may decide which robot/softbot to trust.

Robots with cognitive capacities outside of strictly controlled settings. Safety issues first

Since the introduction of industrial robots in the early 1950s, robots have become a natural part of various manufacturing processes (Nof 1999). Advancements in the fields of electronics, computer science, and mechatronics have made those intelligent tools abundant and robust. A similar revolution is on the way for robots that are designed for non-industrial applications. It is believed that, in the next few decades, intelligent artifacts will be regularly found in private homes as well as in public spaces. Even ambient devices controlling our environments will become intelligent, adaptive and able to communicate.

Of all concerns related to new technologies, safety is the number one priority. Both industrial and social robotics present challenges to human safety. This has been a well-known fact for a long time in industry. To address this issue, industrial robot manufacturers together with end users and other stakeholders have established safety rules and regulations. Unfortunately, the same is not yet true for other domains. In addition, it is not straightforward to transfer the experience gained in industry to address challenges in non-industrial domains. Even the field of industrial robotics is undergoing a major transformation. Today an industrial robot is behind fences in a robot cell as it is hazardous to approach it. In the future, cognitive robots will move freely and work in close interaction with humans.

Robots in close collaboration with humans

The system that has been developed in our group is a typical example (Akan et al. 2010; Çürüklü et al. 2010) of present day advanced intelligent machinery. It is a robot that is controlled by natural language. When required, it gives spoken feedback to the user. The camera mounted above the gripper resembles an eye. The robot detects objects, and more importantly, interprets its environment based on the input from the sensors and the speech commands from the user.

Robots that can reason and talk are an intellectual challenge to humans, especially robots that have a capability to move freely and naturally among humans. Psychologists have already noticed a peculiar attachment of people to even rudimentary intelligent artifacts, such as AIBO or KISMET. With more advanced robots, more involved relationships can be anticipated. Androids, such as ASIMO and Hanson Robotics[1] robots with human-like facial expressions represent the development of robotics in that direction.

Designers of robotic systems need to understand the complex nature of the interaction that will be the result of having this type of system operating in a factory. When the team members can be humans as well as robots, the physical borders between them will become blurred. Moreover, an expert robot may be assigned the leading role in a team. The solutions suggested by a robot may be expected to be the optimal ones based on the best available knowledge. Robots can quickly perform complex logical operations, calculations and assessments and maximize

[1] http://world.honda.com/ASIMO/technology/spec.html, http://hansonrobotics.wordpress.com/about/.

preferred outcomes. They can communicate with other robots and use information from databases and the internet much more efficiently than humans do. The problem is that it is not clear today how the engineers should design the software so that its decisions would be not only effective in reaching given goals, but also ethically sound. The answer to that problem is sought within Artificial Morality (Allen et al. 2005, 2006; Wallach and Allen 2009).

Numerous interesting questions arise when the issue of morally responsible artificial agents is addressed by defining autonomous ethical rules of their behaviour—questions addressed by Moor (2006) within the field of Machine Ethics. Even though the implementation of ethics in machines will result in artifactual ethical behaviour, the Machine Ethics[2] itself is developed and implemented by humans whose ethics are regulated by Engineering Ethics, specifically Computing Ethics/Computer Ethics.

All human stakeholders (robot designers, manufacturers, maintenance personnel and users) have specific responsibilities for safe and proper performance, while robots are expected to successively develop artifactual ethical competence, including the ability to take functional responsibility for their own actions. This will be discussed in subsequent chapters.

The requirement for artifactual ethical behaviour of a robot/softbot

The requirements for safe operation of a robot can be fulfilled by both construction and, in the case of intelligent adaptive robots, by inbuilt, self-regulatory behaviour that makes a robot notice and avoid risks and dangers and learn from experience. The agents with morally significant behaviour should have moral responsibility. In the case of a robot/softbot, it may only be *functional artifactual responsibility.* This is very limited in comparison to corresponding human competence, and comes in varying degrees. The idea of artificial agents that are much simpler than the simplest living organisms and thus much easier to model and simulate is presented in Danielson (1992), and Floridi and Sanders (2004). The main advantage of the results obtained from simple agent models is that they help us to rethink the essentials of ethical conduct and present a basis for the development of ethically responsible artifacts with morality "by design".

Adding the requirement for ethical behaviour to a robot or a softbot does not mean that the artifact should possess the totality of human moral capacities, just as an intelligent artifactual system does not possess all of the human intelligent capabilities. The requirement of artifactual ethical competence for a robot/softbots should be in accordance with the artifactual agent's intelligence and depend on the application. So, for example, softbots trading stocks or cars (Grodzinsky et al. 2011) should have the intrinsic, ethical norms of their particular domain of agency. We want such bots to behave decently and not to cheat us in order to maximize the profit for someone else.

Significant development in the field of social robotics calls for Requirements Engineering for which ethical conduct will be essential. Thus, we have good reasons to study technological development scenarios in which robots/softbots are becoming so sophisticated that they possess different degrees of artifactual morality along with artifactual intelligence.

Moral responsibility, classical versus pragmatic

One of the essential characteristics of ethical behaviour is moral responsibility. Two main approaches to moral responsibility are *the classical approach,* which implies that artifacts cannot be ascribed responsibility, and *the pragmatic approach,* which implies that some artifacts can be ascribed various degrees of artifactual (functional) responsibility.

Classical approaches, against artifactual responsibility

The Stanford Encyclopedia of Philosophy provides the following explanation (McKenna 2009):

> *A person who is a morally responsible agent is not merely a person who is able to do moral right or wrong. Beyond this, she is accountable for her morally significant conduct.* ***Hence, she is an apt target of moral praise or blame,***[3] ***as well as reward or punishment***. (emphasis added)

This is a widely held position, found even in Eshleman (2009), Siponen (2004), and Sullins (2006).

A frequent argument against ascribing moral responsibility to artificial intelligent system holds that it is pointless

[2] See e.g. http://www.aaai.org/Press/Reports/Symposia/Fall/fs-05-06.php AAAI Fall 2005 Symposium on Machine Ethics and http://uhaweb.hartford.edu/anderson/machineethicsconsortium.html. Machine Ethics Consortium.

[3] This understanding of the necessary **connection of responsibility with blame** builds on the underlying supposition that the error always is a problem of an individual agent and not a problem of a system as a whole. It also implies that the system of individual agents is regulated by order and punishment. This is fundamentally different from the modern safety culture approaches that, starting from individual responsibility, emphasize global properties of system safety.

to assign praise or blame as it has *no meaning* to an artificial agent (Floridi and Sanders 2004). In that case, we must reflect on the meaning of *meaning*. An agent may well be programmed so that it has meaning for praise and blame in the same way that it has meaning for goals and obstacles. Going one step further would be building emotions/synthetic emotions into artifacts as discussed in Coeckelbergh (2010), Becker (2006), Arkin (1998), Fellous and Arbib (2005), and Minsky (2006). Emotions appear to be a very powerful regulatory mechanism. How they will be implemented in robots remains to be seen, but the field of synthetic emotions is developing progressively (Vallverdú and Casacuberta 2009).

In order to decide whether an agent is morally responsible for an action, it is often believed to be necessary to consider two parts of the action: *causal responsibility* and *mental state* (Nissenbaum 1994). The mental state aspect of a moral action is what classically distinguishes morally responsible agents. Traditionally, only humans are considered to be capable of moral agency. The basis of the human capability of action is intention (Johnson 2006). Intentionality enables learning from mistakes, regret of wrongs and wish to do right—all of which are seen as typically human abilities.

A frequent argument against the ascription of moral responsibility to artificial intelligent systems is that they do not have the capacity for mental states like intentionality. The problem is that it is unclear what such a mental state entails (Floridi and Sanders 2004). In fact, even for humans, intentionality is ascribed on the basis of observed behaviour, as we have no access to the inner workings of human minds—which is much less than our access to the inner workings of a computing system (Coeckelbergh 2010; Dodig-Crnkovic 2006).

In addition, both arguments above against ascribing moral responsibility to artificial intelligent agents (no mental states and no meaning for blame) (Johnson 2006; Johnson and Miller 2006; Grodzinsky et al. 2008) come from the view that an artificial intelligent agent is primarily an isolated entity. Nevertheless, in order to address the question of moral responsibility, we must view intelligent agents as parts of a larger sociotechnological organization. From that perspective, as already pointed out, responsibilities are distributed and networked in such a complex system and ascribing (a degree of) responsibility to an intelligent agent has essentially a regulatory role.

It should not be forgotten that organizations, such as corporations and similar sociotechnological systems, also have a collective (group) moral responsibility (Floridi and Sanders 2004; Coleman 2008; Silver 2005), which differs from individual human responsibility.

Finally, artificial intelligence is making continuous progress and learning autonomous intelligent agents will successively become so advanced that we will have no problem in ascribing to them intentions (again *artifactual intentions* that come in degrees, on a par with artificial responsibility and artificial agency). Thus, both conditions will be fulfilled: causal responsibility and a "mental state"—intentionality for an artificially intelligent agent.

Pragmatic functional approaches: moral responsibility as a regulatory mechanism

However, questions of intentionality (Dennett 1994) and the free will of an agent are difficult to address in practical engineering circumstances, such as the development and use of intelligent, adaptive robots/softbots. Thus, Dennett and Strawson suggest that we should understand moral responsibility not as an individual duty, but instead as a role that is defined by externalist pragmatic norms of a group (Dennett 1973; Strawson 1974). We adopt this pragmatic approach which is closer to actual praxis and robot applications.

Moral responsibility can best be seen as a social regulatory mechanism that aims at enhancing actions considered to be good, simultaneously minimizing what is considered to be bad. It makes sense for an agent who is able to perform a task and to assess its outcome. That is what Wallach and Allen call the "instrumental approach":

> We take the instrumental approach that while full-blown moral agency may be beyond the current or future technology, there is nevertheless much space between operational morality and "genuine" moral agency. This is the niche we identified as functional morality. (Wallach and Allen 2009)

Along a similar line, Asaro suggests that we view robots as sociotechnical systems (Huff 2010) and therefore think of a continuum of agency between completely amoral and fully moral agents. Robots may be found along this continuum and, as they develop more complex agency, they will be expected to show greater ethical competence.

For Dennett, moral responsibility is a rational and socially efficient policy and it is the result of natural selection within cooperative systems (Dennett 1973; Järvik 2003). Moral responsibility as a regulative mechanism shall not primarily focus on locating the blame but more importantly on assuring future appropriate behaviour of a system.

Moral responsibility may be considered to be the obligation to behave in accordance with an accepted ethical code (Sommerville 2007). It influences the behaviour of agents who have been assigned responsibilities (Dodig-Crnkovic 2005). In Software Engineering practice, for example, moral responsibility is a subfield of system dependability in which practical questions of allocation,

acceptance, recording and discharge of responsibilities are addressed.

Artificial intelligence and artificial morality

Dodig-Crnkovic (2006), Dodig-Crnkovic and Persson (2008), and Adam (2008) all emphasize the similarities of artificial intelligence and artificial morality:

Artificial/artifactual intelligence is defined as an ability of an artificial agent to accomplish tasks that are traditionally thought to require human intelligence.

Artificial/artifactual morality can be defined as an ability of an artificial agent to behave in a way that is traditionally thought to require human morality.

Artifactual intelligence is not the same phenomenon as human intelligence, but it can produce the same specific behaviours. As artifacts become more and more intelligent and autonomous, we expect them to behave in accordance with our value systems and ethical norms.

Responsibilities in a sociotechnological system in practice

Responsibility for a system involving technological artifacts must take into account designers, manufacturers and users, as well as the technological artifacts themselves (Johnson and Powers 2005). It is not only human agents that, by engineering and operating instructions, can influence the morality of artificial agents. Artifacts, as actors in a sociotechnological system can impose limits on human actors and influence them too (Adam 2005; Latour 1992). Despite this, the study of the relationships between humans and technology until now has always emphasized the one-way impact originating in human designers and manufacturers. Nevertheless, when predicting global development, we have to take into account that, while we are changing technology, technology in turn is changing us (Becker 2006; Russell and Norvig 2003).

Production and use of intelligent artifacts have increased the complexity of sociotechnological systems. Even if today's robots/softbots are used mostly as automatic tools without ethical capabilities (Lin et al. 2008), artifacts display more and more autonomous, morally significant behaviour, and the possibility of ascribing moral responsibility to intelligent machines has been discussed (Matthias 2004; Johnson 2006; Floridi and Sanders 2004; Stahl 2004). A system that takes care of certain tasks intelligently, learns from experience and makes autonomous decisions, gives us good reasons to talk about a system as being responsible for a task.[4] No doubt, technology is morally significant for humans, and the responsibility for a task with moral relevance should be accompanied by functional moral responsibility.

> Regardless of whether artificial morality is genuine morality, artificial agents act in ways that have moral consequences. This is not simply to say that they may cause harm—even falling trees do that. Rather, it is to draw attention to the fact that the harms caused by artificial agents may be monitored and regulated by the agents themselves. (Allen et al. 2005)

Regulatory mechanisms in safety critical systems

It is expected that any technology that is subject to significant uncertainty and has a potentially high impact on society will be handled cautiously, and intelligent systems surely fall into this category, where *the precautionary principle* applies (Hansson 1997, 1999; Montague 1998; Som et al. 2004). Thus, responsibility for preventing harm and the burden of proof that it is not harmful is something that manufacturers of intelligent technology are responsible for. According to the precautionary principle, we have not only a right, but also a moral obligation to anticipate the ethical consequences of reasonably foreseeable paths of development.

When it comes to the practical applications, the most important issue is safety. Based on experience with safety critical systems, such as aerospace, transportation and healthcare systems, several levels of organizational and physical barriers are typically necessary to be ready to cope with different degrees of severity of malfunctions. One can say that the sociotechnological structure that supports their functioning consists of safety barriers that prevent and mitigate malfunctions. The central issue is to ensure the safe operation of the system under normal conditions, which is complemented by the preparedness for mitigation of abnormal and accidental conditions (Dodig-Crnkovic 1999).

The macro-level safety assurance must take into account everything from technical issues, issues of management and anticipating use and effects, to larger issues on the level of societal impact (Huff 2004; Asaro 2007). The crucial ethical concerns for engineers are risk identification

[4] Davis (2010) e.g. distinguishes among nine senses of "responsibility", one of those being (e) a responsibility as domain of tasks (things that one is supposed to do) –which is a type of responsibility we argue should be ascribed to robots.

and assessment and the assurance of sufficient safety levels (Shrader-Frechette 2003; Larsson 2004).

Ethical aspects specific to physical safety

The coexistence and interaction of human and artifactual intelligence include issues of physical safety of people involved in interactions with intelligent artifacts. Many robots are specifically constructed to contribute to increased human safety. These include rescue robots or robots operating in human-hostile environments. On the other hand, some robots possess the power to handle huge loads together with a wide range of motion and may pose physical danger to people. Problems of safety are traditionally regulated by safety standards. The new ANSI/RIA/ISO 10218-1:2007 Standard addresses safety requirements for emerging robot technologies, including human-robot collaboration, robot-to-robot synchronization, and vision-based safeguarding systems. In parallel with the development of new types of robots, new standards are issued. Corresponding standards are needed for softbots, cognitive robots and other artificial agents used in social robotics.

Thinking about robotics safety is usually associated with humanoids or automatic mechanical machines while intelligent technologies develop also in the direction of embedded ambient intelligence (Magnani 2007; Floridi 2007). Crutzen underlines the ethical significance of "invisible" ambient intelligence (Crutzen 2006). This type of intelligent control of environments urges us to re-think our fundamental concepts and values, including the *idea of a good life, personal integrity and indeed personhood*. Proactivity is a preferred course of action (Brey 2006) as in the general case of a precautious approach to new technologies. Even when it comes to safety, ambient intelligence is vital as a complex intelligent learning system that controls basic functions, such as temperature and lighting, permission to enter or leave the house, activating mechanisms in the case of emergency, storage and the accessibility of information. Safety aspects of ambient intelligence call for specific safety standards.

Safety by design. Safety culture. High Reliability Organizations

Robotic safety is one of the most fundamental questions on which the future of robotics depends and, according to (Veruggio and Operto 2008), security and reliability are the most important ethical codes of conduct.

> Safety by design is a concept and movement that encourages construction or product designers to "design out" health and safety risks during design development. The concept supports the view that along with quality, programme and cost; safety is determined during the design stage.[5]

The focus of "safety by design" is given to properly designed and constructed engineering solutions as the most important *first line of defense* against potential safety risk. Safety by design is complemented by *the second line of defense,* which involves safety culture; administrative procedures or protective measures to eliminate and reduce risks. Understanding the risks and learning from experience allows one to anticipate probable safety threats and establish preventive measures.

Safety culture is thus a control mechanism on the organizational level with the aim of establishing and sustaining a high level of safety in a sociotechnological system. It can be defined as the shared commitment of management and employees to ensure the safety. Efficient communication and, thus, openness and transparency are the basis of safety culture, which acknowledges the inevitability of error, and proactively seeks to identify potential threats and weaknesses in the organization.

> To establish a culture of safety, an organization must change from one of blame for errors to one where errors are seen as opportunities to learn and improve. A culture of safety recognizes that errors exist and are a part of the business, and deals with them in a non-punitive manner (unless behaviour is truly egregious).[6]

Organizations that are known to be complex and risky, yet safe and effective are known as High Reliability Organizations (HRO:s), and can be found in aviation, the healthcare sector or the nuclear industry.[7] The characteristics of an HRO are systematic reporting, openness and transparency, learning organization and accountability. An HRO maintains a balance between personal responsibility and the responsibility of the entire organization where the individual is protected and procedures are in place for reporting and reviewing events. *"When something goes wrong, the focus is on what, rather than who, is the problem. The intent is to bring process failures and system issues to light, and to solve them in a non-biased manner". (ibid.).*

In High Reliability Organizations, moral responsibility is distributed among moral agents, which are assigned different tasks. Responsibility is firstly connected to the

[5] http://en.wikipedia.org/wiki/Safety_by_design.

[6] http://patientsafetyed.duhs.duke.edu/module_c/what_do_we_mean.html.

[7] The Fukushima disaster reminds us how risky the nuclear industry is and how highly reliable it is under normal conditions. It also gives us reason to think about the consequences of rare catastrophic events.

execution of a task, while incidents, errors, failures, and accidents must first be connected with the learning of an organization to tackle risks properly and to never repeat mistakes.

Today punishment, blame, and prosecutions are by no means the main aspect of safety—neither for humans nor for machines. The paradoxical situation where machine designers are supposed to search for ways to build intelligent agents that can be *punished* is based on a misunderstanding of the role of punishment. *Punishment is not a goal in itself.* If the safety of a system can be established by reprogramming the agent, then that is the solution to the problem. Some would object that punishment is missing because the punishment is seen as compensation for the harm. However, in modern technological systems, work is distributed among many people, processes are complex and it is generally impossible to place the blame on one person in order to punish her/him. The main goal in the present circumstances is to assure the safety of the system, which is achieved in High Reliability Organizations by learning and constantly improving processes and routines.

Fundamental to safety is a broad safety culture with a pro-active attitude and a defense in depth, which effectively contribute to the development of safety and should be taken into account in the macro-level Requirement Engineering for ethical robots.

The rules: moral responsibility of engineers

From the discussion of the different approaches to responsibility (classical vs. pragmatic), it is obvious that at this stage there is no consensus about the necessity of Machine Ethics that would assure the ethical behaviour of robots and softbots. Questions asked are: Is it possible? Is it desirable? Contrary to the arguments put forward in the present article, many would still answer these two questions in the negative. These can be seen as the Requirements Engineering issues. We already mentioned the widespread concern that building moral characteristics, such as responsibility, into artifacts may result in humans handing over all responsibility to the robots/softbots. The mobilization of engineers on the issues connected to intelligent computing artifacts is therefore considered to be central:

"The Rules, Moral Responsibility for Computing Artifacts", an initiative lead by Keith W. Miller https://edocs.uis.edu/kmill2/www/TheRules aims to:

> (…) reaffirm the importance of moral responsibility for (computing) artifacts, and to encourage individuals and institutions to carefully examine their own responsibilities as they produce and use them. (Miller 2011)

Miller has gathered together an international Ad Hoc Committee for Responsible Computing consisting currently of 50 members working on improving drafts, (Miller 2011; Pimple 2011). As an illustration, here is the first rule, of five, according to the present draft 27:

> Rule 1: The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artifact as part of a sociotechnical system.

This is in line with the option of building robots ethical by design.

Artifactual morality by design

As already argued, the claim that artificial agents cannot be assigned responsibility based on the fact that blame and punishment has no meaning for an artificial agent can be met by counter-arguments from the perspective of safety culture.

Firstly, in modern High Reliability Organizations the primary interest is not in assigning the blame, but in learning from experience.

Secondly, in the case of egregious behaviour of an intelligent artificial agent a corrective mechanism equivalent to regret or remorse can be introduced, as previously discussed, either by synthetic emotions (Coeckelbergh 2010) or in some other way that will prevent an artifactually intelligent autonomous robot/softbot from repeating its mistakes.

In addition to the management of individual agents, for the future hybrid sociotechnical systems of humans and intelligent autonomous artifacts (Nobre et al. 2009) it is necessary to develop intelligent learning management of the system as a whole, with an emphasis on constant learning and development of culture of safety and moral responsibility.

> The development of machines with enough intelligence to assess the effects of their actions on sentient beings and act accordingly may ultimately be the most important task faced by the designers of artificially intelligent automata. (Allen et al. 2000)

Artifactual responsibility is parallel to artifactual intelligence. It must be specific for specific types of intelligent agents. It can by no means reduce the responsibility of engineers. Instead, according to the precautionary principle, the engineers are expected to ensure the ethically acceptable behaviour of artificial agents.

We argue that moral responsibility in sociotechnological systems, including autonomous, learning intelligent robots/softbots is best viewed as a regulatory mechanism, and it follows a pragmatic (instrumental, functionalist) line of thought. For all practical purposes, the question of responsibility in learning intelligent systems may be addressed in the same way as safety in traditional safety critical systems in High Reliability Organizations, which constantly confront complexity, risks and unexpected situations, but operate safely and effectively.

The development of autonomous, learning, morally responsible, intelligent, artificial agents must rely on several responsibility loops. These are the awareness and preparedness for handling risks on the side of designers (Davis 2010), manufacturers, implementers, users and maintenance personnel, as well as the support of society at large, which provides a response to the consequences and expectations of the use of technology. This complex system of shared responsibilities should assure safe functioning of a hybrid organization of humans and intelligent machines.

Conclusion

Our conclusion is that artifactual (functional) morality should be built into future robots/softbots, with the aim of ensuring their ethically adequate behaviour. *Artifactual morality should be seen as a necessary companion to artifactual intelligence in artificial agents.* In the similar way that artifactual intelligence is a function of an engineered system that would require intelligence in humans, the artifactual ethical behaviour is a function of an artifact that would require morality in a human agent.

Floridi and Sanders (2004) have proposed a generalization of the notion of *agenthood*, providing the common framework for studying a variety of agents—from the extremely simple, "mindless" ones to complex, cognitive, intelligent systems. We continue in the same direction and propose the notion of *responsibility* to be generalized, encompassing different levels of responsibility so that a certain degree of (functional) responsibility can be ascribed to machines within a techno-social system. This artifactual responsibility can be compared to the responsibility within a hierarchical organization. The leaders of an organization have more (and a more complex) responsibility than the members at the basic level of the hierarchy. Nevertheless, all contributions count and are necessary for an organization to function safely and ethically. Errors and failures are used to learn and improve and not primarily to punish.

Marino and Tamburrini discuss moral responsibility and liability in a case when substantial limitations exist in predicting the behaviour of robots that learn and adapt based on experience:

> One has to take into account the fact that robots and softbots - by combining learning with autonomy, proactivity, reasoning, and planning - can enter cognitive interactions that human beings have not experienced with any other non-human system. (Marino and Tamburrini 2006)

Robots ethical by design based on Requirements Engineering cannot be expected to emerge at once. They will be improved iteratively in an evolutionary process, rather than produced at once as a result of intelligent design, because many of the phenomena in the real world applications are expected to be emergent (Brey 2008).

One thing to keep in mind is that we do not want machines to behave like humans. We want them to behave as *ideal humans*. In the same way that we expect them to calculate without error, which they actually do, while humans err (errare humanum est!), we expect machines to *behave blamelessly*. We obviously still have a long way to go to achieve that goal.

Acknowledgments The authors want to thank Mark Coeckelbergh for the enlightening discussion of several ideas central to this paper—distributed responsibility, artifactual morality and the role of emotional intelligence. We also gratefully acknowledge Keith Miller's insightful response to an earlier version of this article. Last, but not least, we greatly appreciate the three anonymous reviewers' valuable comments.

References

Adam, A. (2005). Delegating and distributing morality: Can we inscribe privacy protection in a machine? *Ethics and Information Technology, 7*, 233–242.

Adam, A. (2008). Ethics for things. *Ethics and Information Technology, 10*(2–3), 149–154.

Akan, B., Çürüklü, B., Spampinato, G., Asplund, L., et al. (2010). Towards Robust Human Robot Collaboration in Industrial Environments. In *Proceedings 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 71–72).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*, 149–155.

Allen, C., Smit, I., & Wallach, W. (2006). Why machine ethics? *IEEE Intelligent Systems*, July/August 2006, pp. 12–17.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–25.

Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge: MIT Press.

Asaro, P. M. (2007). *Robots and responsibility from a legal perspective. Proceedings of the IEEE 2007 International Conference on Robotics and Automation*, Workshop on RoboEthics, Rome.

Aurum, A., & Wohlin, C. (2003). The fundamental nature of requirements engineering activities as a decision-making process. *Information and Software Technology, 45*(14), 945–954.

Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In Patrick Lin, George Bekey, & Keith Abney (Eds.), *Robot*

ethics: The ethical and social implication of robotics. Cambridge, MA: MIT Press.
Becker, B. (2006). Social robots—emotional agents: Some remarks on naturalizing man-machine interaction. *International Review of Information Ethics (IRIE), 6*, 37–45.
Brey, P. (2006). Freedom and privacy in ambient intelligence. *Ethics and Information Technology, 7*(3), 157–166.
Brey, P. (2008). Technological design as an evolutionary process. *Philosophy and Design, 1*, 61–75.
Bynum, T. W., & Rogerson, S. (Eds.). (2004). *Computer ethics and professional responsibility* (pp. 98–106). Kundli, India: Blackwell.
Capurro, R., & Nagenborg, M. (Eds.). (2009). *Ethics and robotics*. Amsterdam: IOS Press.
Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford: Oxford University Press.
Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society, 24*, 188–189.
Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. In: *Ethics and Information Technology*, published on-line ISSN 1388-1957, 18 March 2010.
Coleman, K. G. (2008). Computing and moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition). http://plato.stanford.edu/archives/fall2008/entries/computing-responsibility/.
Crutzen, C. K. M. (2006). Invisibility and the meaning of ambient intelligence. *International Journal of Information Ethics, 6*(006), 52–60. Ethics in Robotics.
Çürüklü, B., Dodig-Crnkovic, G., & Akan, B. (2010). Towards industrial robots with human like moral responsibilities. In *Proceedings 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 85–86).
Danielson, P. (1992). *Artificial morality virtuous robots for virtual games*. London: Routledge.
Davis, M. (2010). Ain't no one here but us social forces: Constructing the professional responsibility of engineers. *Science and Engineering Ethics*, Issn: 1353-3452, pp. 1–22.
Dennett, D. C. (1973). Mechanism and responsibility. In T. Honderich (Ed.), *Essays on freedom of action*. Boston: Routledge & Keegan Paul.
Dennett, D. C. (1994). The myth of original intentionality. In E. Dietrich (Ed.), *Thinking computers and virtual persons: Essays on the intentionality of machines* (pp. 91–107). San Diego, CA and London: Academic Press.
Dodig-Crnkovic, G. (1999). *ABB atom's criticality safety handbook, ICNC'99 Sixth International Conference on Nuclear Criticality Safety*, Versailles, France.
Dodig-Crnkovic, G. (2005). On the importance of teaching professional ethics to computer science students, computing and philosophy conference, E-CAP 2004, Pavia, Italy. In L. Magnani (Ed.), *Computing and philosophy*. Associated International Academic Publishers.
Dodig-Crnkovic, G. (2006). *Professional ethics in computing and intelligent systems. Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*, Espoo, Finland, Oct 25–27.
Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. In A. Holst, P. Kreuger & P. Funk (Eds.), *Tenth Scandinavian Conference on Artificial Intelligence SCAI 2008*. Vol. 173, Frontiers in Artificial Intelligence and Applications.
Edgar, S. L. (1997). *Morality and machines: Perspectives on computer ethics*. Sudbury, MA: Jones and Bartlett Publishers.
Eshleman, A. (2009). Moral responsibility. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 Edition). http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/.
Fellous, J.-M., & Arbib, M. A. (Eds.). (2005). *Who needs emotions?: The brain meets the robot*. Oxford: Oxford University Press.
Floridi, L. (2007). *Distributed morality in multiagent systems*. Paper Presented at CEPE 2007, San Diego. http://cepe2007.sandiego.edu/abstractDetail.asp?ID=40.
Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.
Gates, B. (2007). A robot in every home. *Scientific American, 296*, 58–65.
Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 11*(1), 115–121.
Grodzinsky, F., Miller, K., & Wolf, M. (2011). Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics and Information Technology, 13*(1), 17–27.
Hansson, S. O. (1997). The limits of precaution. *Foundations of Science, 2*, 293–306.
Hansson, S. O. (1999). Adjusting scientific practices to the precautionary principle. *Human and Ecological Risk Assessment, 5*, 909–921.
Huff, C. (2004). Unintentional power in the design of computing systems. In T. W. Bynum & S. Rogerson (Eds.), *Computer ethics and professional responsibility* (pp. 98–106). Kundli, India: Blackwell Publishing.
Huff, C. (2010). "Why a sociotechnical system?" http://computingcases.org/general_tools/sia/socio_tech_system.html.
Järvik, M. (2003). How to understand moral responsibility?, Trames, No. 3, Teaduste Akadeemia Kirjastus, pp. 147–163.
Johnson, D. G. (1994). *Computer ethics*. Upper Saddle River, NJ: Prentice-Hall, Inc.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*, 195–204.
Johnson, D. G., & Miller, K. W. (2006). A dialogue on responsibility, moral agency, and IT systems, *Proceedings of the 2006 ACM symposium on Applied computing table of content* (pp. 272–276). Dijon, France.
Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*, 99–107.
Larsson, M. (2004). *Predicting quality attributes in component-based software systems*, PhD Thesis. Sweden: Mälardalen University Press. ISBN: 91-88834-33-6.
Latour, B. (1992). Where are the missing masses, sociology of a few mundane artefacts, originally. In Wiebe Bijker & John Law (Eds.), *Shaping technology-building society. Studies in sociotechnical change* (pp. 225–259). Cambridge, Mass: MIT Press.
Levy, D. N. L. (2006). *Robots unlimited: Life in a virtual age*. Natick, Massachusetts: A K Peters, Ltd.
Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. http://ethics.calpoly.edu/ONR_report.pdf.
Lin, P., Bekey, G., & Abney, K. (2009). Robots in war: Issues of risk and ethics. In Rafael Capurro & Michael Nagenborg (Eds.), *Ethics and robotics*. Heidelberg, Germany: AKA Verlag/IOS Press.
Magnani, L. (2007). *Distributed morality and technological artifacts. 4th International Conference on Human being in Contemporary Philosophy*, Volgograd. http://volgograd2007.goldenideashome.com/2%20Papers/Magnani%20Lorenzo%20p.pdf.
Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics (IRIE), 6*, 46–51.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.

McKenna, M. (2009). Compatibilism. In: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 Edition). http://plato.stanford.edu/archives/win2009/entries/compatibilism/.

Miller, K. W. (2011). Moral responsibility for computing artifacts: The rules. *IT Professional, 13*(3), 57–59.

Minsky, M. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. NY: Simon & Schuster, Inc.

Mitcham, C. (1995). Computers, information and ethics: A review of issues and literature. *Science and Engineering Ethics, 1*(2), 113–132.

Montague, P. (1998). The precautionary principle, Rachel's environment and health weekly, No. 586. http://www.biotech-info.net/rachels_586.html.

Moor, J. (1985). What is computer ethics? *Metaphilosophy, 16*(4), 266–275.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems, IEEE computer society*, pp. 18–21.

Moravec, H. (1999). *Robot: Mere machine to transcendent mind*. Oxford, New York: Oxford University Press.

Nagenborg, M. (2007). Artificial moral agents: An intercultural perspective. *International Review of Information Ethics, 7*(09), 129–133.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM, 37*(1), 73–80.

Nobre, F. S., Tobias, A. M., & Walker, D. S. (2009). Organizational and technological implications of cognitive machines: Designing future information management systems. *IGI Global*. 1–338. doi: 10.4018/978-1-60566-302-9.

Nof, S. Y. (Ed.). (1999). *Handbook of industrial robotics* (2nd ed.). Hoboken, New Jersey: Wiley.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. *Proceedings of International Conference on Software Engineering (ICSE-2000)* (pp. 4–11). ACM Press: Limerick, Ireland.

Pimple, K. D. (2011). Surrounded by machines. *Communications of the ACM, 54*(3), 29–31.

Russell, S., & Norvig, P. (2003). *Artificial intelligence—A modern approach*. Upper Saddle River, NJ: Pearson Education.

Scheutz, M. (2002). *Computationalism new directions* (pp. 1–223). Cambridge, Mass: MIT Press.

Shrader-Frechette, K. (2003). Technology and ethics. In R. C. Scharff & V. Dusek (Eds.), *Philosophy of technology—The technological condition* (pp. 187–190). Padstow, United Kingdom: Blackwell Publishing.

Silver, D. A. (2005). Strawsonian defense of corporate moral responsibility. *American Philosophical Quarterly, 42*, 279–295.

Siponen, M. (2004). A pragmatic evaluation of the theory of information ethics. *Ethics and Information Technology, 6*(4), 279–290.

Som, C., Hilty, L. M., & Ruddy, T. F. (2004). The precautionary principle in the information society. *Human and Ecological Risk Assessment, 10*(5), 787–799.

Sommerville, I. (2007). Models for responsibility assignment. In G. Dewsbury & J. Dobson (Eds.), *Responsibility and dependable systems*. Kluwer: Springer. ISBN 1846286255.

Stahl, B. C. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines, 14*, 67–83.

Strawson, P. F. (1974). *Freedom and resentment, in freedom and resentment and other essays*. London: Methuen.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*(12), 23–30.

Vallverdú, J., & Casacuberta, D. (2009). Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and artificial intelligence. *IGI Global*. doi: 10.4018/978-1-60566-354-8.

Van de Poel, I. R., & Verbeek, P. P. (Eds.) (2006). Special issue on ethics and engineering design. *Science, Technology and Human Values 31*(3), 223–380.

Verbeek, P.-P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In P. E. Vermaas, P. A. Kroes, A. Light, & S. Moore (Eds.), *Philosophy and design* (pp. 91–103). Berlin, Germany: Springer.

Veruggio, G. (2006). *The EURON Roboethics Roadmap, Humanoids'06*, December 6, 2006, Genoa, Italy.

Veruggio, G., & Operto, F. (2008). *Roboethics, chapter 64 in springer handbook of robotics*. Berlin, Heidelberg: Springer.

Wallach, C., & Allen, W. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Warwick, K. (2009). *Today it's a cute friend. Tomorrow it could be the dominant life form*. Times of London 2009-02-25. http://www.timesonline.co.uk/tol/comment/columnists/guest_contributors/article5798625.ece.