【Articles】

The Cultural Impact on Ethics: Robotic Agency in Socio-Technical Systems

GRÜNEBERG, Patrick

Abstract

Robotic technologies gain increasing impact on the human lifeworld. Usage of robotic agents in fields such as healthcare, medical service or military provokes farreaching ethical issues. In particular, the question of responsibility asks for a thorough ethical assessment of robotic technologies. However, facing this challenge standard ethical approaches provoke a conflict. On the one hand, they conceive robots as individual agents and seek to ascribe moral agency comparably to human agents. In this view, an agent gains ethical impact by moral agency. On the other hand, it can be shown by an analysis of cognitive architectures underlying robotic behavior that moral status ascription does not work for robots so that they cannot be regarded as moral agents. In the consequence, the ethical impact of robots can hardly be recognized within standard ethical approaches.

At the same time, robotic architectures show that robots develop their agency in interaction with humans. In this view, cultural contexts of robotic usage imply a shift from an individualistic to a relational perspective according to which robots gain ethical relevance as parts of socio-technical systems. The implied cultural impact on ethics is exemplified by presenting the seal robot Paro, a healthcare and therapeutic robot which is used in Japan and Germany. By linking cognitive analysis of robotic architectures with the concept of socio-technical systems and its reference to cultural contexts, the relational approach explains why robots do matter ethically.

Keywords: responsibility, cognitive architectures, context-sensitivity, Paro

1 Introduction

With the increasing usage of robotic agents in various fields such as labor and services, military and security, research and education, entertainment, medical and healthcare or personal care and companions, related ethical issues are also on the table. As humans might be injured or commodities be damaged, it is necessary to clarify the central question of responsibility. It has to be figured out and finally specified in legal terms “who pays the bill” if a robot misbehaves. Beside responsibility issues, there are other severe ethical questions regarding

the impact of robots on humans during interaction as in the case of health and care robotics. Or issues concerning an equal distribution and availability of robotic technologies and their impact on employment.(1)

All these issues raise the question of how to assess robotic behavior ethically. Approaching this task in terms of applied ethics, moral theory is applied to robotics comparably to bioethics concerning genetics or environmental ethics concerning human interventions in the natural environment. However, when it comes to robotics there arises a particular conflict: On the one hand, it is for sure that robots do matter ethically as they have begun to invade the human lifeworld. On the other hand, common ethical theory can hardly explain why robots do matter ethically. As will be shown by means of a cognitive analysis of robotic agents, the common procedure of ascribing moral agency to individual agents does not work in the case of robots because robots elude the ascription of properties which characterize a moral agent. Accordingly, their specific type of agency and the related *ethical* impact on humans is not explained sufficiently.

At the same time, cognitive analysis of robotic agency shows a path to resolve this conflict. When shifting the conception of robots from an individualistic to a relational perspective, robots can be considered as parts of socio-technical systems. Whereas standard ethics builds on culturally independent moral criteria of an individual agent, my proposal draws on the dependency of robotic agents on their interactional relations with humans so that their ethical impact can be assessed within cultural contexts of their interaction.(2) Building on the analysis of robotic architectures and linking the ethical assessment to specific interaction scenarios, this investigation builds on the interconnection of scientific, cultural and ethical studies in order to understand the ethical dimension of robotic agency.(3)

In the following, I will present the ethical standard concept of individual agents and moral status ascription in terms of autonomy, moral reasoning and responsiveness (2.1). While robots are able to act autonomously, responsiveness is out of reach of current robotics. Considering cognitive architectures which fail to deal with the frame problem adequately, it can be shown that robotic agents do not implement moral reasoning (2.2). In sum, moral agency cannot be

---

(1) For an overview of ethical and legal issues see (Palmerini et al., 2014), regarding employment (Frey & Osborne, 2013).

(2) Coeckelbergh suggests a similar relational shift, but draws on a phenomenolgical and hermeneutical approach in order to overcome the limitations of individualistic approaches (Coeckelbergh, 2011) (Coeckelbergh, 2014). While Coeckelbergh draws on the social construction of robots as agents, I suggest to develop a relational perspective based on the cognitive architecture of robotic agents.

(3) Such an interconnection to deepen the mutual effects is also advocated by (Wagner, 2009) (Wagner, 2014).

ascribed to robots so that the question of their ethical impact remains unanswered. Leaving the individualistic account behind in favor of a relational approach, it can be derived from a cognitive analysis that robots develop autonomous agency in an interactional loop with humans (3.1). Further considering the dependency of robots on humans (3.2), robots are located within socio-technical systems (3.3). Thereby, the general question of moral responsibility of an individual agent is pinned down to the question of legal liability of all actors of a socio-technical system containing robots so that their ethical impact is tied to specific cultural contexts of interaction. Finally, the relational way how robots develop ethical impact is illustrated by an example of care robotics. The case of the robot seal Paro which is used in Japan and Germany (4.1) illustrates how ethical impact assessment is strongly related to cultural contexts (4.2).

2 Robots as individual agents

With the emergence and dissemination of robotic technologies, robot ethics came up as applied ethics in order to deal with the ethical, legal and societal challenges posed by those technologies. Drawing on ethics as a core disciplines of philosophy, robot ethics has been developed in different ways. Verrugio made foundational attempts to set up robot ethics (Veruggio, 2007) (Tamburrini, 2009). In the meantime three distinct meanings can be identified (Abney, 2014, p. 35) (Veruggio & Abney, 2014, pp. 347–348): First, robot ethics states professional ethics of the roboticists, thus seeks to formulate criteria to which engineers, policy makers and everybody who deals with the development and implementation of robotics should adhere to in order to guarantee that robots fulfill ethical criteria and do not harm humans. Other ethical issues concern legal implications such as problems of liability or the availability of robotic technologies and thus questions of their distribution. Second, robot ethics concerns the programming of robots and the implementation of moral rules in the programming code. The goal is to let the robot behave ethically. Third, robot ethics considers self-conscious robots that engage in ethical reasoning and choose a moral code by themselves. In the following, the second meaning is of interest as the first one concerns roboticists and the third one can be regarded as completely out of range of robotic technologies. As will be explained in the next section, the alleged moral status of robots in the second sense builds on an individualistic account of moral status ascription and thus conceives of the robot as an individual agent.

2.1 Good old fashioned robot ethics: ascribing moral agency in terms of autonomy, moral reasoning and responsiveness

Analog to GOFAI (good old fashioned artificial intelligence) (Haugeland, 1985, p. 112), com-

mon ethical accounts can be regarded as “good old fashioned ethics” in the case of robotic agents. Standard ethical accounts come as universalistic accounts which ascribe moral agency based on specific properties of an agent (Floridi & Sanders, 2004) (Misselhorn, 2013). Such properties are derived from the basic concept of personhood and comprise features such as the capacity to act for reasons, behavioral autonomy, understanding, consciousness or responsibility (Misselhorn, 2013, pp. 44, 48). These properties enter into moral reasoning in order to ascribe *moral* agency to an entity (Coeckelbergh, 2014, p. 63): (1) An entity *e* has property *p*. (2) Any entity that has property *p*, has a moral status *s*. (3) Entity *e* has moral status *s*. As there is no fixed canon which properties define moral agency sufficiently, I will focus on three properties which are discussed regarding robotic agents: autonomy, moral reasoning and responsiveness.

The first candidate, autonomy, regards the real-world behavior of an agent, thus how the robot creates its external behavior and intervenes in the world. Considering the complexity of the concept of autonomy and the problem that this concept is not even defined clearly for humans, I will refer here to a definition that has been formulated for describing artificial agents (Gunderson & Gunderson, 2004). An artificial agent is regarded autonomous based on three criteria: (1) An agent generates options to act, (2) selects between these options, and (3) enforces the execution of the selected option so that its choice is not overridden by another agent. Thus, if an artificial agent’s behavior meets these conditions its behavior is regarded autonomous. Regarding the following criteria, autonomy can be seen as a necessary condition for moral agency because an agent who cannot produce a real-world impact by itself cannot be regarded as an agent at all.

The second candidate, moral reasoning and corresponding decision-making, refers to the generation of options to act. There are numerous ways to decide how to act in a specific situation. Therefore, ethics seeks to distinguish between right and wrong behavior. For this purpose, different ethical systems can be applied (Abney, 2014, pp. 36–39). Following a deontological approach (originally formulated by I. Kant), right behavior depends on acting according to certain maxims (intentions) such as being truthful, honest, benevolent, or not making false promises. That means that all actions should follow one or the other duty. These come as moral rules as everyone as a rational being is supposed to accept these rules so that they have to be followed without exception. Circumstances do not alter cases of their implementation. Also consequences of their implementation do not matter ethically.

Opposed to deontology, utilitarianism (defended by J. Bentham and J. S. Mill) solely focuses on the consequences of actions. Right behavior depends on just one rule, the “Greatest Happiness Principle” (GHP): “One ought always to act so as to maximize the greatest amount of

net happiness (utility) for the largest number of people." The goal of morality is to maximize utility with utility being the sum of the good consequences of an action, minus the sum of the bad consequences. Compared to deontology, utilitarianism is a consequentialist theory which only regards the consequences of an action and not the intentions (maxims) ethically relevant.

The third major ethical approach does not ask "What should I do?", but "What should I be?". In this view, morality stems from the agent's character. The goal of virtue ethics (formulated by Aristotle and recently defended by G. E. M. Anscombe) is the virtuous person, a person that seeks for the Good. Virtues are understood as dispositions to act in a certain way and imply habits such as benevolence, civility, self-control or tolerance. Compared to the previously mentioned approaches, actions are not in themselves good or evil, but it depends on a certain context whether an action is morally good. Thus, if an agent is capable of engaging in moral reasoning and subsequent decision-making based on (one of) these approaches, the agent fulfills one more criterion for being a moral agent.

The third property of moral agency, responsiveness, refers to the ability to be held responsible for one's actions. Considering robots as autonomous agents, the question arises who is responsible for their behavior in the case that humans get hurt or commodities get damaged. The issue of responsibility implies that an agent understands what it is doing—that an agent can at least retrospectively reflect on what it did. Furthermore, an agent should understand the consequences of its actions and that these consequences are to be judged in comprehensive cultural and legal contexts which pose certain constraints on individual behavior and possibly result in sanctions. Thus, if an agent is capable of these reflections and judgments, the agent fulfills one more criterion of moral agency.

According to the above mentioned scheme of moral reasoning, the attribution of moral agency concerns the individual agent. As shall be further argued in the next sections, this focus on individual, basically internal properties of an agent, isolates the agent from its environment and abstains from the specific cultural context in which the robot acts. Whereas such a formal conception of moral agency is due to the conceptual and universalistic approach of ethics in general, this approach hits the wall when it comes to robotic agents and their ethical impact (Coeckelbergh, 2011).

2.2 Robotic architecture and moral agency

Current attempts to artificial moral agents (AMA) mainly focus on the first and second property, autonomy and moral reasoning, in order to implement a moral program code (robot ethics in the second sense). The first property, autonomy, can be technically implemented to various degrees (Bekey, 2005, Chapter 1). Even if robotic behavior is usually less complex than

human behavior, robots instantiate autonomous behavior in their respective task domains such as drones controlling their flight route or social robots adapting their behavior to their environment. Regarding moral reasoning, C. Allen and W. Wallach compiled different accounts on artificial moral agency and presented a comprehensive framework for implementing AMAs (Wallach & Allen, 2008) (Allen & Wallach, 2014). Thereby, the programmer's challenge consists in implementing ethical rules in the program code in order to generate moral behavior.

However, technically feasible types of robotic autonomy and procedures for moral reasoning do not imply responsiveness so that the robotic agent could be aware of and account for its behavior. The agent is indeed supposed to behave in ways that are morally acceptable for humans, but it is blind for its behavior.[3] According to the third property, responsiveness, one might object that moral agency implies awareness of one's moral reasoning and that an agent that is blind towards its own behavior cannot be regarded an agent (Himma, 2009). But such an awareness implies a minimum degree of self-awareness and reflective abilities, thus agency in the third sense of robot ethics which is out of range of any technical implementation so that we have to content with "blind" agents in the second sense of robot ethics.

In the following, I will briefly sketch how the major moral theories bear on robot ethics and robotic agency, and exemplify some difficulties of "moral programming". The cognitive analysis of robotic architectures will state that—even if robots act autonomously—current[4] robotic agents do not fulfill the cognitive and behavioral requirements which are necessary for implementing any kind of moral agency. It follows that individualistic moral accounts do not account for the ethical impact of robotic agents on humans and their lifeworld.

（3） Cognitive systems generally fall under the under caveat of being "blind systems" in the sense that identification of relevant data, organization, and finally assignment of meaning to data in order to create behavior completely depends on the programmer so that the system itself does not bring forward any meaningful behavior (Winograd, 1986). Attempts to "unblind" a system aim at coupling it to meaningful contexts of (inter)action so that a system has to generate meaning by shaping its conception of reality and its behavior autonomously (Grüneberg & Suzuki, 2014). As shall be argued in section 3, the underlying concept of this type of relational autonomy cannot any more be conceived as an internal property, but implies an interactional context.

（4） Thereby, I consider moral agency corresponding to the current and near-by state of technology. Of course, there may one day be more sophisticated agents. But the discussion of such an utopistic setup of robots as moral agents comparable to humans might distract from urgent challenges of current robotics. Beside, it remains also questionable whether we want to replicate human conflicts by constructing full-fledged AMAs.

（5） The frame problem originally emerged as a logical problem within classical AI (McCarthy & Hayes, 1969) and has later been applied to more general problems of explaining how cognitive agents gain meaning of the world (Shanahan, 2009).

2.2.1 The frame problem

All the three major moral approaches face the same problem that prevents robots from engaging into moral reasoning: the *frame problem*.[5] The problem builds on the fundamental question what a cognitive agent needs to know in order to come to a decision. How can an agent decide which information in a given situation is relevant, which is irrelevant? Whereas humans can rely on a complex body of experience and common sense knowledge, artificial agents do not have those resources. This problem can be clarified by a simple situation: People meet in a room of a private house, converse and have snacks. If a human, e.g. the child of the host, enters the room, she usually would have no difficulties to understand—or at least to guess— the situation, to distinguish between a birthday party or a funeral, and most important, to act (communicate) accordingly. But how could a robot for home service understand the situation? It poses substantial difficulties to define the kind of perceptual input that is sufficient in order to make the robot distinguish between the different events. The difficulty for the programmer is to anticipate different scenarios and to define distinctive sensual criteria beforehand. Considering changing environments and context-sensitivity, this task requires an impossible computational load for robotic decision-making (Abney, 2014, p. 45).

2.2.2 Moral program codes in top-down architectures

The frame problem regarding "moral programming" is caused by so-called *top-down* architectures.[6] Also labeled as deliberative or cognitivist, these architectures build on the sense-think-act scheme: First, the robot senses the world through its sensors and adds this information to its world model given by the programmer. Here, "world" depends on the sensory equipment and the robot's task domain so that it usually comes as a rather abstract world compared to the human lifeworld. Most importantly, the rich cultural contexts of human lifeworlds are reduced to abstract representations of single aspects so that a robot acts rather independently of the cultural context of its task domain. Second, the robot "thinks", i.e. processes the input and possibly updates the world model, engages in reasoning and action planning based on the world model and some predefined intentions (goals of actions). Third, the robot generates behavior by moving its actuators (such as its hand or wheels) in order to manipulate its environment according to its action plan.

---

(6) Basically, there are three types of cognitive architectures for robots: top-down, bottom-up and their combination (hybrid architectures) (Vernon, Metta, & Sandini, 2007). In the following, I focus on the first type as moral programming seeks to implement explicit moral rules which are implemented by top-down architectures. However, the fact that moral programming basically fails within top-down approaches does not mean that other architectures do better as the frame problem can also be identified in bottom-up architectures (Grüneberg & Suzuki, 2014, pp. 6–7).

The decisive aspect of this architecture is its top-down, i.e. hierarchical structure according to which commands move from top to down layers: after receiving sensory input from the sensory module, the top 'thinking' module (deliberation and planning) controls the bottom motor module (action) so that all behavior is decided *unidirectionally* from top to down. The top layer contains predefined information about the goals of action, the environment (world model) and possible ways of action so that it is here where—theoretically—a moral code could be implemented.

For the purpose of implementing a moral program code, the deontological approach is of particular interest as deontology offers abstract rules whose observance is supposed to guarantee morally adequate behavior. The often cited example here are Asimov's three laws of robotics (Asimov, 1968):

1. "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. "A robot must obey the orders given it by human beings except where such orders would 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 Laws."

However intuitive these rules might appear at a first glance, Asimov himself further specified these rules in order to tackle the frame problem. One deontological version of the frame problem raised by Asimov's laws consists in the assessment of a given situation in order to determine the degree of acceptable risk for humans (cf. law 1). But even from a legal viewpoint, let alone from a moral viewpoint, it is impossible to define corresponding behavioral rules beforehand. A clear example for the infeasibility of an AMA comes with military robots and the distinction between combatants and non-combatants as there does not exist a set of well-defined criteria in order to distinguish between these groups of humans (Sharkey, 2014, p. 118–123). Or take the case of autonomous cars which raise severe legal and ethical conflicts as an appropriate moral behavior cannot be programmed in advance. This would, for example, mean that the programmer instructs the car to decide whether to hit a truck or a pedestrian in case of a conflict—a decision that can hardly be decided *in abstracto*.

The other theoretical option to implement a moral program code is to make use of the utilitarian principle of the greatest happiness.[7] However, alone the estimation of the possible effects of a certain behavior on the condition of the affected humans exceeds any computational power. The utilitarian version of the frame problem for robotic agents comes down to a cal-

(7) Virtue ethics are not further considered as they require some sense of self-awareness which is not feasible for robots.

culational impasse (Abney, 2014, p. 44). As every concrete real-world scenario most probably contain unforeseen variables, the programmer is not able to anticipate every consequence of the robot's action and therefore to decide to what extent an action affects the condition (happiness) of the humans that make part of the scenario.

Considering the frame problem of current robotic architectures, it can thus be argued that robotic processes of internal reasoning might follow some moral principles, but that they fail to apprehend the complexity of cultural contexts of the agent's interaction. Therefore, top-down architectures incapacitate robotic agents from meaningful moral reasoning so that moral agency cannot be ascribed and their ethical impact remains a problem in individualistic approaches.(8)

3 Robots in the loop: culture and context come into play

Following the cognitive implementation of robots which excludes them from being moral agents, the questions remain how robots do matter ethically and how to assess their obvious impact on humans and our lifeworld. In order to understand their ethical impact, I will leave the standard individualistic account of ethics behind and argue for a relational approach to make sense of robotic behavior. For this purpose, robotic agents are located within socio-technical systems, within specific cultural contexts so that their ethical impact can be related to the human lifeworld and is not solely bound to internal properties of an agent.

3.1 Autonomy relies on world-coupling

Regarding the three properties which supposedly constitute moral agency, it can be said from a cognitive viewpoint that autonomy lies at the ground of moral agency: an agent who cannot act autonomously, is not able to engage in moral reasoning or reflection about its actions. Thus, even the comparably narrow concept of robotic autonomy (cf. section 2.1) plays a crucial role in the ethical apprehension of robotic agents. Whereas standard top-down approaches assume autonomy to be an internal feature of the agent's thinking-module, approaches to autonomous mental development suggest that an agent develops its autonomous behavior in interaction with its environment (Weng et al., 2001) (Weng, 2007). The basic idea is that an agent is dependent on environmental feedback and interaction in order to develop its au-

(8) I do not share the optimism of Misselhorn who points to embodied and connectionist systems in order to overcome the frame problem (Misselhorn, 2013, p. 50). As can be shown in the case of subsymbolic metarepresentation (Grüneberg, 2013, Chapter 6.1), such systems are not even able to distinguish between themselves and the world—a caveat that is even admitted by a proponent of embodied architectures (Dreyfus, 2009, n. 102).

tonomous skills. Accordingly, autonomy relies on world-coupling between the robotic agent and its environment, in particular its human interaction partners. In turn, world-coupling broadens the cognitive analysis of moral agency from internal properties to the specific action scenario of the robotic agent and therefore to the cultural context of this interaction.

The relation between the robot's autonomy and its interaction with humans can be exemplified by subjective computing (Grüneberg & Suzuki, 2014). Experiments in socially guided machine learning showed that a robotic agent develops its autonomous skills by means of feedback of a human trainer. These experiments build on robotic agents that are equipped with a specific ability and that can control their respective ability autonomously. In one case, an agent came as a robotic arm whose task consists in balancing an inverted pendulum (Grüneberg & Suzuki, 2014, pp. 10–11). The agent can control various degrees of freedom of the arm so that it is—in principle—capable of balancing the pendulum. However, the agent does not know, i.e. has no pre-programmed behavioral patterns how to fulfill its task. It just starts with random movements. Thus, although it can be said that the agent acts autonomously according to the criteria given in section 2.1, the question arises how the agent can make meaningful use of its autonomous skill and learn to balance the pendulum. Without this behavioral success, the agent's skills could rarely be called a behavior (goal-directed action) as opposed to mere random motion.

Additionally to its ability to control the arm, the robot can receive and interpret feedback of a human interaction partner. While the robot tries to balance the pendulum, a human trainer gives positive and negative feedback which the robot interprets in order to adjust its behavior. Experimental results show that the agent develops its autonomous behavior control and learns to balance the pendulum. Even more interesting is the fact that the robot develops individual learning behavior depending on the human trainer. Thus, the experiment showed how autonomous behavior evolves in interaction with a human trainer.

These results are in line with other approaches (Pfeifer & Scheier, 1999) that ground artificial and therefore robotic intelligence on autonomous skills and their development by means of real-world interaction. From this viewpoint, the ascription of agency refers to an agent in the loop and not solely to internal properties. This loop brings in cultural contexts of the interaction scenario because the interaction between robot and humans depends on the specific habits, attitudes and circumstances of the human counterpart and her lifeworld. This means in the context of moral agency that robotic agents matter ethically due to their autonomous interaction with human counterparts even if their cognitive architectures do not allow ascribing moral agency.

3.2 Asymmetric relation between robots and humans

Another aspect that is usually overlooked while ascribing moral agency based on internal properties is the relation between robots and humans. The standard model assumes a concept of moral agency that is derived from an analysis of human agency. But there is an ambiguity in the meaning of robot ethics: human ethical behavior and robot (ethical) behavior are never on a par. The relation between the two behaviors is never symmetric as robots will not be AMAs in the third sense of robot ethics. Thus, characteristics of human personhood can hardly be ascribed to robotic agents. On the contrary, the relation is always asymmetric as current robots are fully dependent on their human designer, programmer and users (Asaro, 2006, pp. 10–11). The ethical view shifts from individualistic robotic agents to human behavior and human interaction with robots. There is no *individual* robotic agent (Coeckelbergh, 2011, p. 248). As has been argued for by considering the cognitive implementation of robotic autonomy in the loop, also the human-robot relation confirms that robots do not come as individual agents. Instead, they have to be conceived as depending on humans in terms of their very existence and their development of behavior.

3.3 Robots in socio-technical systems: pinning down moral responsibility to legal liability

Following the cognitive analysis of autonomy and the dependency of robotic on human agents, the former ones are cognitively relevant only in interaction scenarios with the latter ones. This *interactional embedding of robots in cultural contexts* leads Asaro to suggest a different definition of robot ethics which concerns (Asaro, 2006, p. 10):

1. Human action through or with robots.
2. Design of ethical robots (i.e. robots that follow certain ethical rules).
3. Ethical relationships between humans and robots.

Contrary to the definition of robot ethics by Veruggio, all these senses have to be addressed simultaneously as they concern the fundamental issue of "how moral responsibility should be distributed in socio-technical contexts involving robots, and how the behavior of people and robots ought to be regulated" (Asaro, 2006, p. 10). Robot ethics should not focus exclusively on roboticists or robots and their properties, but on their interactional relations. For this purpose, he places robots in socio-technical systems which were "established to stress the reciprocal interrelationship between humans and machines" (Ropohl, 1999). Thus, robots appear in their specific contexts of their cultural implementation. Even if robots do not meet the conditions for moral agency, it is nevertheless possible to explain their ethical impact: *robots matter ethically because their autonomous behavior develops during interaction with human*

counterparts and thus affects the latter directly.

This shift from an individualistic to a relational perspective has direct consequences for the ethical assessment of robots. The basic ethical question of responsibility becomes an issue of legal liability (Asaro, 2006, pp. 12–13). It might appear to be too hasty to dismiss ethical criteria completely as legal decisions in turn build on ethical judgments. But for the time being, i.e. to understand the interactional relations of robotic agents, the legal perspective serves to overcome the limitations of individualistic moral status ascription. Instead of formulating abstract criteria for moral reasoning or responsiveness which in turn serve to clarify who is responsible as a *moral* agent (while it remains problematic to decide about the exact consequences of misbehavior for robots), legal liability leads to the question who pays the bill if something goes wrong with a robot. The clarification of the "who" and the "wrong" binds robotic behavior to a concrete cultural context. A robot's ethical impact can be assessed by analyzing the role of all the actors in a socio-technical system ranging from the designer to producers, managers, overseers, policy makers and users. In this view, the robot is never held responsible as an AMA[9], but seen as an extension of human action.[10]

Summarizing the results of the cognitive analysis of robotic autonomy and their agency in socio-technical systems, it can be said that the ethical impact of robots involves the cultural setting of their deployment. Thereby, the relational approach does not exclude ethical concerns completely, but narrows down the discussion of moral agency to concrete scenarios in order to understand the ethical challenges raised by robots: First, not only internal properties of robotic agents, but interactional relations matter. Second, interaction implies context sensitivity so that cultural constraints become an integral aspect of the ethical assessment of robots. Robot ethics shifts away from an individualistic account of moral status ascription to a relational perspective of robots in socio-technical systems. Third, the general question of responsibility is pinned down to the question of the legal impact of robots in specific interaction scenarios (while moral theory still plays an important role in the course of defining criteria of

(9) There are attempts to consider to what extent a robot could be held responsible fot its actions. Sparrow argues that this is not possible as robots cannot be punished because they cannot suffer (Sparrow, 2007). Lokhorst and van den Hoven object to this line of argument as they do not regard punishment as the most effective means for behavioral adjustment of robots (Lokhorst & van den Hoven, 2014). Instead, they suggest "treatment" (149) in terms of a repair or reprogramming of malfunctioning robots as these means are more appropriate ones to reach the goal of correct functioning. In this view, the general moral issue of guilt (as a consequence of violating responsibility) and subsequent punishment shifts to a rather concrete problem of technical modification which involves the mentioned actors in the socio-technial system.

(10) Following this view, procedures of technology assessment can be applied to specify technological, economical, legal, ethical and psychological issues of robotic usage (Decker, 2012) (Decker, 2014).

legal judgment).

4 Context-dependency and ethical assessment

The results of the cognitive analysis of robotic behavior and related problems of moral status ascription will finally be exemplified by a concrete robotic application.

4.1 Care robotics: Paro

Healthcare is one of the fields in which robotic applications are supposed to contribute significantly to the future stabilization of welfare systems. As Japan and Germany are expected to suffer from a shortcoming of workforce, hopes are set on robotic applications in order to support human healthcare personnel. While there are several assistive robotic systems under development, a comparably simple, but effective system has already been implemented and is widely used in Japan and Germany (beside other countries such as Denmark and the US): *Paro*. Paro, presented to the public in 2001, is a "seal type mental commit robot"[(11)] in the shape of a robot baby harp seal that—similar to animal-assisted therapy—is used as a therapeutic tool in order to interact with humans and to create feelings of emotional attachment to the robot.

As an autonomous robot, Paro can move its body parts. It is equipped with five kinds of sensors ranging from tactile to light, audition, temperature and posture sensors so that it reacts to light and dark environments, distinguishes between being stroked or beaten, and being held. It recognizes the direction of voice and certain words such as its name, greetings and praise. Paro also has a learning function and repeats behavior which the human user rewards by stroking or stops behavior which the human user rejects by beating it. Finally, Paro has a soft fur which provides a tactile stimulation close to a real animal.[(12)] Moreover, every single Paro is hand-made and therefore shows some individual variations in its outer appearance.

The developers specify three types of desired effects: "psychological, such as relaxation and motivation, physiological, such as improvement in vital signs, and social effects such as instigating communication among patients and caregivers."[(13)] Research in Japan as well as in the U.S. and several European countries suggests that these effects have indeed been achieved (Kazuyoshi Wada, Shibata, Saito, Sakamoto, & Tanie, 2005) (K. Wada, Shibata, Musha, & Kimura, 2008) (Shibata & Wada, 2011) (Klein, 2011). For the purpose of studying the psychological, physiological and social effects of interacting with Paro, elderly persons residing in nursing

(11) See http://paro.jp/english/index.html (retrieved on 2015-9-21).
(12) See http://paro.jp/english/function.html (retrieved on 2015-9-21).
(13) See http://paro.jp/english/about.html (retrieved on 2015-9-21).

homes were observed during their interaction and later interviewed. Further analysis included video recording and physiological tests such as urine samples. Regarding psychological well-being, Paro made patients laugh, attenuated feelings of loneliness and increased feelings of relaxation. Considerably aggressive patients tended to calm down during the interaction. On a physiological level, neuronal effects on elders affected by dementia and an improvement of resident's vital organs to stress were observed. Finally, social interaction improved significantly when Paro was used in a group session. This aspect has to be emphasized as it does not follow automatically from an application of a therapeutic device that residents leave their isolation behind and engage in group activity (Klein, 2011). There have also been reported some problems with patients who fixated on an individual relation with Paro and cases of non-interest (Klein, 2011). However, the overall implementation of Paro suggests significant positive effects for the elderly resident's of nursing homes and consequently a relief of stress for the care personnel.

4.2 Paro's ethical impact in the socio-technical system of healthcare of elderly persons in nursing homes

Regarding Paro's cognitive architecture, it is obvious that Paro does not count as a moral agent in the standard model. Paro implements a certain degree of behavioral autonomy, but no properties such as moral reasoning or responsiveness. However, considered as part of the socio-technical system of healthcare of elderly persons in nursing homes Paro has a significant ethical impact on humans. In the course of evaluating elderly persons interacting with Paro, several ethical issues are raised (Misselhorn, Pompe, & Stapleton, 2013) (Calo, Hunt-Bull, Lewis, & Metzler, 2011): social interaction between residents, single resident's autonomy, dignity and self-respect. These issues imply the general question of who is responsible for the residents' well-being.

The predominant purport of ethical evaluation tends to the opinion that the advantages of Paro's application outweigh possible ethical disadvantages.[14] For example, Misselhorn objects the claim that interaction with Paro is undignified because the attribution of feelings to Paro is a kind of fraud (Misselhorn et al., 2013, p. 128). Indeed, the cognitive architecture of Paro does not allow concluding that Paro has feelings comparably to animals or humans. It merely reacts to certain external stimuli and thereby provokes its human interaction partner to ascribe feelings to it. However, residents often acknowledge this functionality and distinguish between Paro as a robot and their projection of feelings while still enjoying the interaction with Paro.

(14) For critical views on the utilization of robotic agents in healthcare see (Turkle, 2011, Chapter 6) (von Stösser, 2011).

Even if one acknowledges that the resident's experience of Paro depends on an emotional projection that is not justified by Paro's "inner life", the positive therapeutic effects and the resident's awareness of the situation are regarded to outweigh concerns of dignity. This weighing pattern underlies the discussion of legal and ethical issues regarding Paro.

However, this positive assessment of Paro does not imply an unlimited use of Paro. Klein suggests procedural principles for utilizing Paro (Klein, Gaedt, & Cook, 2013). Furthermore, proponents of Paro emphasize that the cultural context should be considered (Shibata & Wada, 2011, pp. 384–385) (Misselhorn et al., 2013, p. 131) (Klein et al., 2013) as the individual condition of each resident interacting with Paro as well as the specific cultural background finally decide about whether the utilization of Paro achieves the desired effects. For example, research on the subjective evaluation of Paro by visitors of exhibitions showing Paro in Japan, the UK, Sweden, Italy, South Korea, Brunei and the US revealed cross-cultural differences (Shibata, Wada, Ikeda, & Sabanovic, 2009). European visitors tended to accept Paro as a therapeutic tool while Asian visitors[15] tended to accept Paro more as a companion than as a therapeutic tool. To the present day, detailed data of a cross-cultural comparison of elderly residents utilizing Paro does not exist. But the cultural differences of the subjective evaluation of Paro suggest that its impact depends on its interactional contexts. Accordingly, even if Paro does not come as an individual moral agent it gains ethical impact through interaction so that not only its legal but olso its ethical assessment and finally clearance are highly context-dependent.

5 Conclusion

Whereas it is from a cultural studies viewpoint not surprising that the ethical impact of a robotic device depends on its cultural setting, its being part of a socio-technical system, this view is not natural from a good old fashioned ethical viewpoint. In the latter view, ethical impact depends on moral agency which is ascribed to an agent based on certain internal properties. In contrast, it is demonstrated by analyzing robotic architectures that autonomous robotic agents gain ethical impact through their interaction with humans while they do not come as moral agents. Thus, the gap between the failed ascription of moral agency to robots in the standard approach and their factual ethical impact can be overcome by conceiving robotic agents as parts of socio-technical systems. However, it must be noted that moral theory is not dismissed or reduced to cultural (or legal) aspects of specific lifeworlds because moral con-

(15) It must be noted that classification such as "European" and "Asian" are rather coarse. Cross-cultural analysis of the evaluation of humanoid robots between Japan, Korea and the US also revealed significant differences between Japanese and Korean students (Nomura et al., 2007, p. 285).

cepts still play a crucial normative role as criteria for legal judgment. My direction of inquiry rather concerns the significance of moral concepts regarding robotic agency so that I advocate a shift from an individualistic to a relational view on the cognitive implementation of robotic agents in and by socio-technical systems. In this view, cognitive architectures join with cultural and legal constraints in order to understand the ethical impact of robotic technologies.

References

Abney, K. (2014). Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 35–52). Cambridge, Mass.: The MIT Press.

Allen, C., & Wallach, W. (2014). Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 55–68). Cambridge, Mass.: The MIT Press.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics*, *6*(12), 9–16.

Asimov, I. (1968). Runaround [1942]. In *I, Robot* (pp. 33–51). London: Grafton Books.

Bekey, G. A. (2005). *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge Mass. : The MIT Press.

Calo, C. J., Hunt-Bull, N., Lewis, L., & Metzler, T. (2011). Ethical Implications of Using the Paro Robot. In *2011 AAAI Workshop (WS-2011-2012)* (pp. 20–24).

Coeckelbergh, M. (2011). Is Ethics of Robotics about Robots? Philosophy of Robotics Beyond Realism and Individualism. *Law, Innovation and Technology*, *3*(2), 241–250.

Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, *27*(1), 61–77.

Decker, M. (2012). Service robots in the mirror of reflective research. *Poiesis & Praxis*, *9*(3-4), 181–200.

Decker, M. (2014). Who is taking over? Technology assessment of autonomous (service) robots. In M. Funk & B. Irrgang (Eds.), *Robotics in Germany and Japan. Philosophical and technical perspectives* (pp. 91–110). Frankfurt a. M.: Peter Lang.

Dreyfus, H. L. (2009). How Representational Cognitivism Failed and is being replaced by Body/World Coupling. In K. Leidlmair (Ed.), *After Cognitivism* (pp. 39–73). Dordrecht: Springer Netherlands.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds Mach.*, *14*(3), 349–379.

Frey, C. B., & Osborne, M. A. (2013). *The future of employment: how susceptible are jobs to computerisation*. Oxford: Oxford Martin School, Programme on the Impacts of Future Technology.

Grüneberg, P. (2013). *Projektives Bewusstsein. Th. Metzingers Selbstmodelltheorie und J.G. Fichtes Wissenschaftslehre*. Münster: mentis.

Grüneberg, P., & Suzuki, K. (2014). An Approach to Subjective Computing: a Robot that Learns from Interaction with Humans. *IEEE Transactions on Autonomous Mental Development*, *6*(1), 5–18.

Gunderson, J., & Gunderson, L. (2004). Intelligence ≠ autonomy ≠ capability. In E. R. Messina & A. M. Meystel (Eds.), *Performance Metrics for Intelligent Systems: Proceedings of PerMIS '04 Workshop* (pp. 1–7). Gaithersburg, MD: NIST.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: Massachusetts Institute of Technology.

Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties

Must an Artificial Agent Have to Be a Moral Agent? *Ethics and Information Technology*, *11*, 19–29.

Klein, B. (2011). Anwendungsfelder der emotionalen Robotik. Erste Ergebnisse aus Lehrforschungsprojekten an der Fachhochschule Frankfurt am Main. In JDZB (Ed.), *Mensch-Roboter-Interaktion aus interkultureller Perspektive. Japan und Deutschland im Vergleich. Veröffentlichungen des Japanisch-Deutschen Zentrums Berlin* (Vol. 62, pp. 147–162). Berlin.

Klein, B., Gaedt, L., & Cook, G. (2013). Emotional Robots: Principles and Experiences with Paro in Denmark, Germany, and the UK. *GeroPsych*, *26*(2), 89–99.

Lokhorst, G.-J., & van den Hoven, J. (2014). Responsibility for Military Robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 145–156). Cambridge, Mass.: The MIT Press.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, *4*, 463–502.

Misselhorn, C. (2013). Robots as Moral Agents? In F. Rövekamp & F. Bosse (Eds.), *Ethics in science and society: German and Japanese views* (pp. 42–56). München: Iudicium.

Misselhorn, C., Pompe, U., & Stapleton, M. (2013). Ethical Considerations Regarding the Use of Social Robots in the Fourth Age. *GeroPsych*, *26*(2), 121–133.

Nomura, T., Suzuki, T., Kanda, T., Han, J., Shin, N., Burke, J., & Kato, K. (2008). What People Assume about Humanoid and Animal-type Robots: Cross-Cultural Analysis between Japan, Korea, and the USA. Int J. Human. Robot, 5(l), 25–46.

Palmerini, E., Azzarri, F., Battaglia, F., Bertolini, A., Carnevale, A., Carpaneto, J., et al (2014). *RoboLaw. Guidelines on Regulating Robotics*. Regulating Emerging Robotic Technologies in Europe: Robots facing Law and Ethics (report).

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: The MIT Press.

Ropohl, G. (1999). Philosophy of Socio-Technical Systems. *Society for Philosophy and Technology*, *4*(3).

Shanahan, M. (2009). The Frame Problem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009).

Sharkey, N. (2014): Killing Made Easy: From Joysticks to Politics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 111–128). Cambridge, Mass.: The MIT Press.

Shibata, T., & Wada, K. (2011). Robot Therapy: A New Approach for Mental Healthcare of the Elderly – A Mini-Review. *Gerontology*, *57*, 378–386.

Shibata, T., Wada, K., Ikeda, Y., & Sabanovic, S. (2009). Cross-Cultural Studies on Subjective Evaluation of a Seal Robot. *Advanced Robotics*, *23*(4), 443–458.

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

Tamburrini, G. (2009). Robot ethics: A view from the philosophy of science. In R. Capurro & M. Nagenborg (Eds.), *Ethics and Robotics*, (pp. 11–22). Amsterdam: ISO Press.

Turkle, S. (2011). *Alone together: Why We Expect More From Technology and Less from Each Other*. New York: Basic Books.

Vernon, D., Metta, G., & Sandini, G. (2007). A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. *IEEE Transactions on Evolutionary Computation*, *11*(2), 151–180.

Veruggio, G. (2007). *EURON Roboethics Roadmap*. European Robotics Research Network.

Veruggio, G., & Abney, K. (2014). Roboethics: The Applied Ethics for a New Science. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 347–363). Cambridge, Mass.: The MIT Press.

von Stösser, A. (2011). Roboter als Lösung für den Pflegenotstand? Ethische Fragen. *Archiv für Wissenschaft und Praxis der sozialen Arbeit, 3*, 99–107.

Wada, K., Shibata, T., Musha, T., & Kimura, S. (2008). Robot therapy for elders affected by dementia. *IEEE Engineering in Medicine and Biology Magazine*, *27*(4), 53–60.

Wada, K., Shibata, T., Saito, T., Sakamoto, K., & Tanie, K. (2005). Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. (pp. 2785–2790). IEEE.

Wagner, C. (2009): 'The Japanese way of Robotics': interacting 'naturally' with robots as a national character? In: *Proceedings of the 18th IEEE lnternational symposium on Robots and Human lnteractive Communications* (pp.510–515).

Wagner, C. (2014): *Robotopia Nipponica: Recherchen zur Akzeptanz von Robotern in Japan*. Marburg: Tectum.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford, New york: Oxford University Press.

Weng, J. (2007). On developmental mental architectures. *Neurocomputing*, *70*(13-15), 2303–2323.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Autonomous Mental Development by Robots and Animals. *Science*, *291*(5504), 599–600.

Winograd, T. (1986). *Understanding computers and cognition: a new foundation for design*. Norwood, NJ: Ablex Publ. Corp.