What Social Robots Can and Should Do
J. Seibt et al. (Eds.)
IOS Press, 2016

doi:10.3233/978-1-61499-708-5-145

Robots, Autonomy, and Responsibility

Raul HAKLI [a,1] and Pekka MÄKELÄ [a]

[a] *TINT, Department of Political and Economic Studies, University of Helsinki, Finland*

Abstract. We study whether robots can satisfy the conditions for agents fit to be held responsible in a normative sense, with a focus on autonomy and self-control. An analogy between robots and human groups enables us to modify arguments concerning collective responsibility for studying questions of robot responsibility. On the basis of Alfred R. Mele's history-sensitive account of autonomy and responsibility it can be argued that even if robots were to have all the capacities usually required of moral agency, their history as products of engineering would undermine their autonomy and thus responsibility.

Keywords. social robots, responsibility, autonomy, self-control, groups, collective responsibility

1. Introduction

The question in the conference title "What social robots can and should do?" invites one to discuss both responsible robotics and the responsibility of robots. The first concerns responsible ways to deal with robots, e.g. what should we make and let robots do? Obviously robots should not make humans futile and unemployed, but if that happens, the responsibility is not to be ascribed to robots but to humans instead. It seems plausible that robots should not do everything they can, e.g., they should not kill people. Responsible robotics is an issue of collective responsibility of humans. We think that indeed the closer the social robots are to full-blown agency, e.g. the more human-like they are, the issue of responsible robotics becomes ever more important; we should ponder thoroughly whether we want to create artifacts that possibly are monsters like us. Responsible robotics is obviously related to the other leg of our distinction, e.g., if robots are moral agents, responsible robotics is likely to look very different from what it would be if robots were only tool-like instruments.

The other leg of the distinction is responsibility of robots, and this is where we will focus in this paper. Here a natural point of departure is to think of the conditions of moral or legal responsibility and see whether robots do or can satisfy such conditions. Relevant questions include the following: Are robots agents, and if they are, of what kind? How should we compare human and robot agency? Do robots have self-control? What do we mean by autonomous robots?

[1]Corresponding Author: Raul Hakli, Department of Political and Economic Studies, P.O. Box 24 (Unioninkatu 40A), FI-00014, University of Helsinki, Finland; E-mail: raul.hakli@helsinki.fi.

2. Autonomy in Humans and Artificial Agents

Agency is typically attributed to robots in artificial intelligence and in robotics, and there are also several philosophical accounts of agency according to which robots qualify as agents, but the question of autonomy is more difficult. Researchers in AI and robotics standardly talk about autonomous agents in the context of both software agents and physical robots, but it is important to note that they use the term "autonomous" differently from the philosophical usage [1]. In their usage, which is largely adopted in media and everyday parlance, e.g. in such terms as "autonomous vehicles", autonomy means ability to perceive the environment, learn from experience, and act independently of an external operator or controller. According to such an understanding, what makes an agent autonomous is its capability of perceiving and acting in its environment and learning from its experiences in order to modify its behaviour to promote its survival and improve its performance in achieving its tasks [2,3]. Such a conception of autonomy is not relevant for attributions of moral responsibility which in philosophical literature typically requires in addition something like capacity to choose one's goals and act freely (see, e.g., [1]).

In the meta-ethical literature something (see, e.g. [4, p. 3]) along the following lines is meant by "autonomy": etymologically from "auto" (self) and "nomos" (rule or law) "self-rule" or "self-government", and more substantially this involves some kind of attitudes and control over principles, values, desires, and the like one possesses. For instance, Gerald Dworkin [5, p. 108] claims that "autonomy is a second-order capacity to reflect critically upon one's first-order preferences and desires, and the ability either to identify with these or to change them in light of higher-order preferences and values". Here it is useful to follow Joel Feinberg and distinguish among the four meanings of "autonomy" as applied to individual persons: "It can refer either to the capacity to govern oneself... or to the actual condition of self-government ... or to an ideal of character derived from that conception; or ... to the sovereign authority to govern oneself" [6, p. 28].

Autonomy seems closely related to another concept, self-control, which can be understood, in an Aristotelean fashion, as the contrary of akrasia, which positively and more substantially can be characterised as a trait of character exhibited in behaviour that conforms with one's best or better judgement in the face of the temptation to act to the contrary [4].

Both autonomy and self-control are commonly taken to be relevant notions in the characterisation of moral agency that is arguably presupposed by the agentive sense of responsibility.

It is interesting to evaluate robots in light of such conditions independently of the question whether human beings can satisfy such conditions. Indeed we try to avoid any strong pre-commitments as to the issue of the capability of robots to bear normative responsibility: At the start we keep it as a possibility that robots are responsible in the normative sense and yet humans are not. Indeed, in the literature there has been discussion of the possibility of robots outperforming us in moral behaviour for instance in warfare due to their reasoning capabilities and lack of emotional distraction (for critical discussion, see [7]).

However, the standard view is that adult human beings are moral agents responsible of their actions, whereas non-human agents, like animals, robots, and collectives com-

posed of human individuals, are not. We think it is informative to compare the agency of robots on the one hand to animal agency and group agency and to human child's agency and full-blown human agency on the other. This is because such comparisons provide us with different arguments, pro and con, concerning the capacity of different kinds of agents to bear normative responsibility. Although questions concerning attribution of agency or responsibility to robots have often been studied in relation to similar attributions to non-human animals [8], in certain respects it is more fruitful to study them in relation to questions concerning group agency and collective responsibility. This is because animals differ from adult human beings (paradigmatically assumed to be fit to be held responsible) in precisely those respects that are taken to warrant attributions of moral agency, namely features related to rationality, whereas groups are similar in those respects. On the other hand, adult human beings and animals share several common features related to sentience, and hence studying robots in analogy with animals easily leads to considerations of patiency, which (while interesting in its own right) does not get us far in understanding responsibility, which we take to be an agentive notion.

Moreover, robots and collective agents are similar in the sense that they are in certain respects cognitively more powerful than individual human beings. Both robots and groups of individuals can outperform individual human beings in some cognitive tasks due to their larger computational or memory resources. Finally, assuming that it is possible to consider robots and groups as agents, they are both artificial agents in the sense that they are both constructed by human beings.

This last common feature in particular suggests that the analogy between robots and groups might enable us to modify existing arguments concerning group responsibility to the case of robots. In the case of groups, it can be argued that even if they could in principle be fit to be held morally responsible, their agency depends in a constitutive way on the agency of their makers, hence it would be unfair to hold them responsible in a normative sense [9].

At the core of such line of argumentation is the worry concerning the autonomy of collective agents. Here psychological autonomy is taken to amount roughly to the autonomy regarding various aspects of one's mental or psychological life including one's pro-attitudes. One reason why groups can be considered as agents is that their attitudes are psychologically discontinuous with their members' attitudes, that is, their actions are not a function of the attitudes of the individual members [10]. The worry stems from the fact that collective agents are agents created or designed by the individual members to serve their purpose(s), and agents that depend on the individual members in executing their agency. As the autonomy is a necessary condition of an agents' fitness to be held responsible, the capacity of collective agents to bear responsibility is also undermined by the same token as their autonomy is.

The crux of the reasoning here is that the psychological discontinuity is part of the rational design of the collective agent made by the individual members. To put the point bluntly, collective agents considered as separate agents are under the control of the individual members in having the attitudes that are not a function of the attitudes of the individuals because that is exactly what the individual members do and should want the collective agents to have. This sort of control seems to question the autonomy of collective agents, and thus the capacity of such agents to bear responsibility in their own right, despite the psychological discontinuity between the individual members and the collective agent.

The individual members are in control of the core of the value-set, preferences, and the capacity of critical reflection attributable to the collective agent in the following sense. The collective agent is identified or "identifies itself" with the purpose which is the purpose of the individual members, qua individuals, and which provides the individuals with a reason to create a collective agent. That purpose becomes the core of the value-set attributable to the collective agent. The decision-making procedure which is constitutive to the collective agent qua autonomous agent can also be described as the constitutive element of the capacity of critical reflection attributable to the collective agent. The shared purpose of the individual members is not only the reason for the existence of the collective agent, but the purpose also provides the collective agent with a perspective or orientation from which its faculty of critical reflection operates.

Collective agents can also be argued to lack the kind of control over their actions that is required by moral responsibility. Assume that the individual human members of a collective agent retain their intentional agency and control over their own actions as they become members of a collective agent. Assume also that a collective agent can act only via the intentional actions of its individual human members. The intentional actions of a collective agent are constituted by the intentional part actions qua part actions of the individual members of the collective agent. The individual members, qua agents distinct from the collective agent, are in reason-sensitive control over their part actions qua part actions; they are in control over the actions they knowingly perform as actions required by the collective agent's action.

In light of these assumptions, it seems that even in the case of a continent action by a collective agent the ultimate responsibility lies at the level of individual members.

On the basis of the reasoning summarily characterised above one can construct an argument building on widely shared assumptions concerning the ontological and "psychological" nature of collective agents to the effect that collective agents are dependent and subordinate agents in a way that makes it unfair to hold them morally responsible in their own right. (See [9] for the full argument.)

The argument does not fully carry over to the case of robots, however. Even though the agency of robots depends in a constitutive way on the agency of their makers in the sense that their capacities to have intentional states and to act result from the design and implementation of these features by human engineers, at some point they gain autonomy in the engineering sense and become capable of acting without continuous human control. This is not the case with human groups, which always remain under the control of their members. However, this is not necessarily a dead-end. There may be grounds to argue that the fact that the robots' capacities are engineered suffices to undermine their responsibility. Let us first look at which kinds of capacities have often been required of moral agency and moral responsibility.

3. Properties Required for Moral Agency and Responsibility

It is a rather common and firm intuition that robots are not moral agents—they are not autonomous and morally responsible agents. However, it is only an intuition and intuitions do not carry much evidential weight. In what follows we try and offer some aspects of agency that may offer elements for a diagnosis of the differences between human beings and robots that may account for the intuition. This is not to say that they solve the

problem whether robots are moral agents, solving the problem would require an account of necessary and sufficient conditions of moral agency which we are not in the position to offer. We can, however, ponder whether some of the aspects of the diagnosis are such that they could be argued to be necessary conditions of a morally responsible agent and whether it is reasonable to claim that robots can satisfy them.

Another common and firm intuition is that we, adult human beings, are moral agents. Our self-image as moral agents consists in part of the understanding of ourselves as both biologically and psychologically extremely complex creatures with various capacities falling into the following classes: (1) Intentionality, (2) Rationality and action, (3) Sentience, (4) Autonomy, (5) Normative understanding, and (6) Sociality and personhood (see Table 1).

Table 1. Capacities and attributes suggested as necessary for moral agency.

Capacities related to					
Intentionality	Rationality and action	Sentience	Autonomy	Normative understanding	Sociality and personhood
believing	reasoning	self-awareness	setting goals	recognizing normative reasons	social commitment
desiring	action and omission	consciousness	self-control	awareness of responsibility	communication
intending	decision-making and planning	emotions	weakness of will	moral reasoning	reciprocity
self-reflection: higher-order int. states	deliberation	empathy	critical reflection of values	rule-following	recognition

As already mentioned, the list above is not providing us with sufficient nor necessary conditions of moral agency. However, it is a list of features that we commonly think can be attributed to a regular human adult. In light of a diagnosis it is of some interest to see how a robot, the most sophisticated robot imaginable, perhaps one that has evolved as a result of an artificial evolutionary process, would fair in terms of these features.

With respect to capacities related to intentionality, there is some controversy, but probably most roboticists and a large number of philosophers are willing to grant at least first-order intentional states to robots. They would admit that robots can have beliefs, desires (understood as pro-attitudes) and intentions.

If first-order intentionality is granted, then some capacities related to rationality and action seem to be rather universally accepted, and robots can consequently be taken to satisfy conditions of agency. Robots are taken to be capable of goal-directed action. It can be said that robots can make decisions, plan their actions, and that robots are capable of omission as well.

Higher-order intentionality, capacity for deliberation and setting goals are more debatable as are capacities related to sentience, normative understanding and personhood. It is imaginable that a son of a strong willed soldier becomes a soldier himself, due to his upbringing he comes to adopt many beliefs, desires and values from his tyrant-like father. The son starts killing people for his living. He becomes a mercenary. However, after a couple years of blood shedding, he starts feeling uncomfortable with the job, he

starts reflecting on the aims of the army, his personal goals, and values. He realises he has done many bad things in his life. He goes through a serious period of an identity crisis and ends up giving up pretty much all the values and commitments he has held so far. He becomes a peace activist. He aims at redemption. He becomes a morally good person. This very simple story is not implausible when told about a human being, but would it be even imaginable if the protagonist were a robot? Can a robot set goals, aims, and ends to itself and change them? (Consider a robot thinking "I used to be a robot working in a factory but got bored with it and now I am doing maths.")

There are many other perplexing questions concerning potential capacities of robots. Can a robot exercise self-reflection on its values directing its goal setting, planning and actions? ("I used to think that war is right but then I realised it only causes suffering and now I understand causing unnecessary suffering to human beings and other robots is wrong, I am all for peace and love now?") Can a robot be aware of its responsibility? Can a robot recognise normative reasons? ("I want to do some more maths, but there is a kid buried to the back yard of my neighbour, perhaps I should go and save her?") Perhaps a robot can suffer from the weakness of will? Is a robot capable of having various kinds of motivational urges that may cause trouble for rational action? If a robot is capable of rational action, then maybe also incontinent action (even though designing robots riddled with weakness of will and incontinent action may not be in anyone's interest)?

We tend to think of robots as designed, developed, and engineered by someone else, mechanical, target bound, having narrow and specific intentional horizon, lacking feelings and emotions, etc. However, they are also capable of learning, have enormous computational power, they can perform extremely complex tasks, they are perhaps more reliable than human beings (e.g., self driving cars). What are the relevant aspects for the evaluation of their moral citizenship or lack thereof? Maybe the moral agency is an issue of degree? Would it be wrong to treat robots as second rate members of the moral community? Why would it be important to be fair to a robot? These are vexing questions that we cannot try to answer here.

What we can do is to note a common line of argumentation in many attempts to show that robots cannot bear moral responsibility: Take one such property or capacity X on focus, suggest that it is necessary for moral responsibility, and then argue that robots fail to instantiate X. Searle's classic Chinese Room Argument [11] against robot intentionality was not originally targeted against robot responsibility, but has later been appealed to in that context as well [12,13]. Roboticists and philosophers inspired by Dennett's idea of intentional stance grant intentional states like beliefs, desires and intentions to robots, but not necessarily higher-order intentional states that are required for self-reflection. Dennett himself takes higher-order intentionality as a necessary precondition for moral responsibility and sees little evidence for such a capability in present-day robots [14].

Raffaele Rodogno similarly argues that current robots are not moral agents. He builds his argument on a neo-sentimentalist view of morality, according to which grasp of morality requires both feelings and a capacity to make normative attribution of emotions at issue [13]. Capability of having feelings does not suffice, but an understanding of when certain attributions are appropriate is required as well. Understanding the meaning of moral wrong is in part constituted by our having certain justified moral sentiments. Robert Sparrow argues that robots lack the capacity to suffer and this makes it difficult to hold them responsible because holding someone responsible requires at least a conceptual possibility to reward and punish them [15]. Similarly, Christian Neuhäuser finds

several capacities, like an evaluative system for considering actions from a moral point of view, for which it is doubtful whether robots can be said to have them [16].

In general, this is a good argumentative strategy. However, pointing to a capacity *X* that current robots lack, easily leads to the roboticists' reply that they can engineer *X* in the next generation of robots. And indeed they have claimed of virtually all the capacities mentioned above that these can be implemented in robots.

4. Argument against Robot Responsibility

Our argumentative line here is different in the sense that we are willing to grant, for the sake of the argument, that all of these capacities can be implemented in robots. So let us assume that we eventually manage to build robots that have higher-order intentionality, consciousness, and emotions. They are able to critically reflect on their values, adopt those values (or "identify" with those values) that they take to be best supported by reasons, and be aware of the consequences of their actions. The question we then ask is this: Suppose we manage to implement all the capacities in robots that are conceptually required for moral agency, does this guarantee artificial moral agency? Does this give us robots that are morally responsible for their actions?

Our reply is: Not necessarily! We claim that having these capacities may not be enough, because it matters how these capacities were acquired. Here we lean on Alfred R. Mele's history-sensitive externalism about autonomy and responsibility [4, pp. 144–176]. According to it, there is more to being autonomous and responsible for one's actions over a stretch of time than what goes on inside an agent during that time. Whether agents are autonomous depends on the agents' causal histories, how they came to possess the intentional attitudes, values, and capacities that they currently have. Mele gives several examples of two "psychological twins" whose relevant mental capacities are identical, but one has *authentically* acquired those mental capacities and is for that reason autonomous whereas the other is not, because her capacities were acquired by external manipulation. Consider the following example [4, p. 145]:

> Ann is an autonomous agent and an exceptionally industrious philosopher. She puts in twelve solid hours a day, seven days a week; and she enjoys almost every minute of it. Beth, an equally talented colleague, values a great many things above philosophy, for reasons that she has refined and endorsed on the basis of careful critical reflection over many years. She identifies with and enjoys her own way of life—one which, she is confident, has a breadth, depth, and richness that long days in the office would destroy. Their dean (who will remain nameless) wants Beth to be like Ann. Normal modes of persuasion having failed, he decides to circumvent Beth's agency. Without the knowledge of either philosopher, he hires a team of psychologists to determine what makes Ann tick and a team of new-wave brainwashers to make Beth like Ann. The psychologists decide that Ann's peculiar hierarchy of values accounts for her productivity, and the brainwashers instill the same hierarchy in Beth while eradicating all competing values—via new-wave brainwashing, of course. Beth is now, in the relevant respect, a "psychological twin" of Ann. She is an industrious philosopher who thoroughly enjoys and highly values her philosophical work. Indeed, it turns out—largely as a result of Beth's new hierarchy of values—that what-

> ever upshot Ann's critical reflection about her own values and priorities would have, the same is true of critical reflection by Beth. Her critical reflection, like Ann's, fully supports her new style of life.

According to Mele, although Ann is autonomous, Beth is not, even though they are psychologically similar in all relevant respects. If this is correct, it shows that autonomy depends on history.

The moral of Mele's story to robots is obvious: Our suggestion here is that even though robots could be programmed and engineered to have all the capacities required of moral agents, they would still not be moral agents, because such programming and engineering is closer to the kind of autonomy-undermining manipulation that deprives the authenticity of their capacities than to the authentic acquisition characteristic of the capacities of real moral agents. It is precisely the fact that the responsibility-relevant property X was engineered that undermines the responsibility attribution to the agent: Robots cannot be morally responsible because they are designed and programmed by other agents to have the "character" they have.

It should be noted that our point is not just that robots are programmed, hence not free, hence not autonomous. That point has been made several times in the literature, and it has a standard objection: One could argue that pretty much the same applies to the "paradigmatic moral agent", that is, adult human beings. Human beings are what they are because of their genetic programming (DNA), upbringing, and contingent influences from their environment.

True, it can well be the case that human beings' characters are "determined" by external factors and influences. However, from the point of view of autonomy relevant for moral agency, the point is not that being externally determined undermines autonomy, and *mutatis mutandis*, moral agency. According to history-sensitive externalism the relevant issue is the kind of determination, and here one could try and argue that robots by practical necessity have their character in virtue of such a process, that they are necessarily manipulated and thus they are not autonomous because of their history. Moral responsibility requires autonomous agency in the sense that S is a morally responsible agent only if S is an autonomous agent (at least in some measure).

Let us look in a bit closer detail at Mele's conception of authenticity. According to him, [4, p. 166], a "necessary condition of an agent S's *authentically* possessing a pro-attitude P (e.g. value or preference) that he has over an interval t is that it be false that S's having P over that interval is, as I will say, *compelled**—where compulsion* is compulsion *not arranged by S*."

Compulsion of a pro-attitude, according to Mele, requires that the shedding, that is, eradicating or significantly attenuating, of the pro-attitude is beyond the agent's control in the sense that her psychological constitution precludes her from voluntarily producing conditions that would empower her to shed the pro-attitude. Such a pro-attitude is *practically unsheddable* [4, p. 153].

Even though Mele's focus is on cases where the agent previously possesses authentic values which are then bypassed by some kind of manipulation, this is not necessary for his idea of compulsion [4, p. 168]:

> Suppose that it is logically possible for a devil to create an agent with certain unsheddable pro-attitudes, "identifications," and reasons for identification already in place at the time of creation. In such a case, the devil did not "bypass"

> the agent's capacities for control over his mental life in producing these items, for the agent had no such capacities when the items were produced. Still, the agent is reasonably viewed as being possessed of some *compelled* pro-attitudes with which he identifies (for reasons); call them compelled *innate* pro-attitudes. [...] This notion of "innate" pro-attitudes may be extended to apply to agents who, after coming into existence, but prior to having any (relevant) capacities for control over their mental lives, are subjected to pro-attitude engineering.

This seems to be the case with robots: Their pro-attitudes are engineered by the designers and implementers, and the robots are not able to shed those pre-programmed pro-attitudes (as our imaginary soldier above seemed to be). Of course, not all of their pro-attitudes are pre-programmed. They, too, can learn from the environment, adapt, and change their behaviour by various machine learning techniques, such as reinforcement learning [17], leading them to behave in ways unpredictable even to their designers. However, it does not seem to matter whether the robot learns or is completely pre-programmed, because some pro-attitudes will have to be programmed in either case. Otherwise the robot lacks motivation to do anything, including motivation to learn. It will not do anything! It has to have some kind of objective, like a reward function to maximize as in reinforcement learning, in order for it to do anything. Taking the intentional stance, such a function is to be understood as a pro-attitude. Such pro-attitudes guide the robot's learning and, as a consequence, its future behaviour, and they seem to be practically unsheddable. Because such pro-attitudes are programmed by other agents, they are compelled in Mele's terminology. Hence, they are not autonomously acquired and lack authenticity. As a consequence, the responsibility of the robot's actions that are rooted in such pro-attitudes is undermined.

Let us still note that it is not essential to the argument that the robots are programmed by other intentional agents. Mele's conception of compulsion does not require that the source of it is an agent [4, p. 169]. Bluntly, the argument is that the "character" of robot agents is a result of the kind of manipulation which undermines autonomy and responsibility. This is at the core of justification of our firm intuition that robots are not moral agents.

5. Conclusions

Above we have discussed, the possibility of full-blown citizenship of robots in our moral community. This would require that we take them to be moral agents, which in turn requires autonomy and moral responsibility. Rather than solving the problem whether robots can be morally responsible, our aim has been more diagnostic: We have tried to find support for the common intuition that robots are not autonomous and morally responsible in the same sense in which human beings are taken to be.

Another approach and a topic to be studied in future work is to consider another important part of moral agency, namely the awareness of the moral dimension of our lives. That awareness is significant for us because of our ever present weaknesses, we have to struggle to do the right thing, we have urges to act against our better knowledge, we have uncertainty about what is right and what is wrong, we are painfully aware of our responsibility for our decisions, actions, and choices. One might claim that the significant role of morality in our lives is anchored in part in our weakness. It may be that robots do

not come across as having weaknesses, indeed, it may be that they are too strong, have too much computational power and are too indefatigable and consistent to be moral in the full sense. Could it be that their moral agency is problematic, not because they do not qualify, because they lack some capacities, but because their capacities are flawless to an amoralizing extent?

Acknowledgements

We thank Søren Andersen for comments and corrections. This research has been supported by the Academy of Finland through the funding of the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences (TINT).

References

[1] Willem F.G. Haselager. Robotics, philosophy and the problems of autonomy. *Pragmatics and Cognition* **13** (2005), 515–532.
[2] Cynthia L. Breazeal. *Designing Sociable Robots*. MIT Press, 2002.
[3] Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach, third edition*. Pearson Education, 2010.
[4] Alfred R. Mele. *Autonomous Agents: From self-control to autonomy*. Oxford University Press, 1995.
[5] Gerald Dworkin. *The Theory and Practice of Autonomy*. Cambridge University Press, 1988.
[6] Joel Feinberg. *Harm to Self*. Oxford University Press, 1986.
[7] John P. Sullins. Robowarfare: Can robots be more ethical than humans on the battlefield? *Ethics and Information Technology* **12** (2010), 263–275.
[8] David J. Gunkel. *The Machine Question: Critical perspectives on AI, robots, and ethics*. MIT Press, 2012.
[9] Pekka Mäkelä. Collective agents and moral responsibility. *Journal of Social Philosophy* **38** (2007), 456–468.
[10] Philip Pettit. Groups with minds of their own. In: Schmitt, F. (ed.), *Socializing Metaphysics*, Lanham, Md.: Rowman and Littlefield (2003), 167–193.
[11] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences* **3** (1980), 417–424.
[12] Kenneth E. Himma. 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** (2009), 19–29.
[13] Raffaele Rodogno. Robots and the limits of morality. In: Nørskov, M. (ed.), *Social Robots: Boundaries, potential, challenges*. Ashgate (2016), 39–55.
[14] Daniel C. Dennett. When HAL kills, who's to blame? Computer ethics. In: Stork, D. G. (ed.), *HAL's Legacy: 2001's computer as dream and reality*, Cambridge, MA: MIT Press (1997), 351–365.
[15] Robert Sparrow. Killer robots. *Journal of Applied Philosophy* **24** (2007), 62–77.
[16] Christian Neuhäuser. Some sceptical remarks regarding robot responsibility and a way forward. In: Misselhorn, C. (ed.), *Collective Action and Cooperation in Natural and Artificial Systems: Explanation, implementation and simulation*, Springer (2015), 131–146.
[17] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 1998.