Ethics in Progress (ISSN 2084-9257). Vol. 10 (2019). No. 2,
Art. #9, pp. 98-111. DOI:10.14746/eip.2019.2.9

Moral Competence and Moral Orientation in Robots

André Schmiljun

(Humboldt-Universität zu Berlin; schmiljun@insystems.de)
ORCID: 0000-0002-9415-8495

1. Introduction

"What is moral competence and why do we have to promote it?" (Nowak 2016, 322). Ewa Nowak asks this question in her review of Georg Lind's book (2016) *How to Teach Morality. Promoting Deliberation and Discussion, Reducing Violence and Deceit*. She concludes – while agreeing with Lind's definition – that moral competence is an important ability "to solve problems and conflicts on the basis of universal moral principles through thinking and discussion, instead of using violence, deceit, and force" (Lind 2016, 45). Furthermore, moral competence is decisive for a "competent moral subject and interactive member of a society" (Nowak 2016, 330). This is true particularly in the case of democracies where individuals need to behave morally and respect other beings in a "process of participation" (Steć 2017, 44). It is therefore not surprising that even in robot ethics – where it is still a rather new term – moral competence is seen as a crucial feature for "social robots" (Malle 2014, 189). Any robot – as Malle says – becomes a "social robot" the moment it starts to "collaborate with, looks after, or helps humans" (Malle 2014, 189). As robots take over more and more application domains in our society, being embedded as assistive robots in health care, companion robots for children and elderly people at home or as combat robots in military (Lin 2014, 11), these machines are no longer only objects (moral patients) to us, but turn into subjects of interaction and communication and act as moral agents.[1] Thus, one major aim of robot ethics from the very beginning (Misselhorn 2018, 90), is to define rules and principles that we can integrate into robots (Malle 2016, 243). Shall robots act, for instance, on Isaac Asimov's *Three Laws of Robotics*? Shall they

1 Sullin defines three requirements for robotic moral agency: autonomy, intentionality and responsibility: "Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities" (Sullin 2006, 29).

behave according to a consequentialist, deontological or virtue-based theory? As Nowak asks, "What kind of ethics should be implemented in AI and what kinds of competencies should be experientially acquired by AI? Should it be more sophisticated or practicable ethics?" (Nowak 2017, 187). Most importantly, it must be ensured that robots will not be dangerous or cause "psychological and social problems" (Veruggio *et. al*. 2011, 22) or evoke "harm to humans and other entities worthy of moral consideration" (Wallach 2010, 243). Autonomous robots must be capable to cope with our social and moral norms in society that are "deeply ingrained in human cognition and behavior". If they fail to comply with them it could result into different social reactions, from blame to rebuke to "full-fledged legal consequences" (Scheutz 2017, 57).

Two major strategies (the top-down and bottom-up strategies) are currently discussed in robot ethics for moral integration. I will argue that both strategies are insufficient. Instead, I agree with Bertram F. Malle and Matthias Scheutz that robots need to have moral competence. However, I argue that we should not define moral competence merely as a result of different "elements" (Malle 2016, 245) or "components" (Scheutz 2017, 61). My suggestion is to take Lind's *Dual Aspect Dual Layer Theory of moral self as a framework for moral robots.* Lind provides not only another perspective and vocabulary for the discussion, but also highlights that moral competence is only one aspect of moral behavior. The second aspect is moral orientation.[2] As a result, the thesis of this paper is that integrating morality into robots has to include moral orientation and moral competence. Scheutz and Malle focus primarily on the cognitive aspect of behavior (moral competence). Scheutz understands moral competence - similarly to Lind – as an "ability to judge situations based on moral principles such as norms and values and make morally and ethically sound decisions" (Scheutz 2017, 57).

The paper is structured as follows: I will start with a brief description of the two strategies for moral integration, before turning to the concept of moral competence in robot ethics in the final section.

2. What Kind of Ethics Is the Right One for Robots?

It is indeed a challenge to define what kind of ethics should be implemented into robots that collaborate closely with us every day. In literature, this issue appears usually in general theoretical debates, though there are a few experiments testing different ethics in computer programs (for example *Jeremy* – a prototype of computational program developed by Anderson, Anderson on the basis of Bentham's Utilitarianism; or *W. D*., based on Rawls' philosophy and W. D. Ross's seven prima facie duties, see McLaren 2018, 300). It is assumed that robots might become more than just "tools" and instruments

2 In his book, Lind doesn't address the question of whether intelligent machines could be as morally competent as human beings. Although Lind doesn't deal with this issue, his theory delivers a valid and – in psychology – widely accepted concept that can be useful for robot ethics.

used by "human beings" (Andersons 2018, 1). Considering robots as moral patients, robot ethics focuses on our behaviour and attitude, such as concern, respect, care or moral responsibilities that we as human beings owe to them (Schmiljun 2017, 71). But the moment robots will be able to act as moral agents, it evokes the question if these machines should be equipped with the same moral obligations and responsibilities as us, including the additional same political rights and laws we share for living. Thus, Hall asks provocatively if "a computer was as smart as a person, was able to hold long conversations that really convinced you that it understood what you were saying, could read, explain, and compose poetry and music (…) – would it be murder to turn it off" (Hall 2018, 32). In short, what requirements must be fulfilled for us to accept a robot to be an equal member of our society?

Robots need to follow some ethical principles or must be constructed in a way that makes their actions predictable or at least comprehensible for us. Basically, there are two considered strategies of how to implement morality into robots: the top-down or bottom-up strategy. Both strategies are linked with different approaches in computing software and ethical theories (Misselhorn 2018, 96). Whereas bottom-up strategies are rather connected with metaethics, moral particularism or virtue-based theories, top-down implementations contain general ethical principles. Probably the most infamous rule-based top-down approach are the *Three Laws of Robotics* formulated by Isaac Asimov in "Runaround" in 1942. If a robot obeys the laws it is understood to be "morally correct". V*ice versa* if it doesn't follow the rules it "acts immorally" (Abney 2014, 36).

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.

The problem with these three laws is that they don't fit with the complexity of our reality. Imagine we let an autonomous car run on Asimov's rules. If the car has to decide whether to injure itself including its occupants or some other uninvolved people on the street it will be confronted with a heavy dilemma which can't be solved in the given rule framework. Somehow, one law or another has to be broken. As Jason Miller puts it, "Designing ethics settings for collision management can be ethically challenging since many hard cases are often highly personal" (Millar 2017, 24). Thus, Gips argues that Asimov's laws are not "suitable"; they are "laws for slaves" (Gips 2018, 244), representing a "Sklavenmoral" (Misselhorn 2018, 111). "We want our robots to behave more like equals, more like ethical people" (Gips 2018, 244).

Another classic example of top-down approaches are consequentialist theories

with its most popular scheme called the utilitarianism. The idea of consequentialist theories is to judge actions by their consequences. In utilitarianism proposed by Bentham in the late 18th century, an action in order to be moral must produce the "greatest balance of pleasure over pain". If we want to know whether an action is good or bad we need to take a look at its results and sum up the "pleasure or pain for each person" (Gips 2018, 245). The trouble with this theory is again its inability to capture the complexity of our reality and uncertainty with regard to a definition. Is the greatest pleasure for human beings, for example, also the greatest pleasure for animals or ecosystems? Who shall be taken into calculus when making decisions? Shall a robot judge every human as equal no matter what certain preferences, lust or suffer they have. I agree with Misselhorn that this approach is simply not practical or helpful for a cognitive system to make decisions (Misselhorn 2018, 98) as it always has to take an uncountable number of facts into consideration before performing any action. In fact, the robot would be overwhelmed and paralyzed by processing possible options.

A third possibility of a top-down strategy are deontological ethics. In contrast to consequentialist theories, a deontological ethic judges not the consequences of an action but the actions in and of themselves. Independent of the expected result, it is the intrinsic motivation of an action that makes it moral or immoral. If an action follows a universal principle it can be evaluated as morally correct. Imagine again the former example: if an autonomous car is given the universal principle »Don't kill« and is involved in an accident where it unintentionally kills people, it can still be judged as morally good because the car acted according to its universal principle and saved the lives of its occupants.

A prominent model of this ethical strategy is the Kantian categorical imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction." (Kant 1993/1785[1], 30) Recently, Nowak published a version of a categorical imperative-based ethic where she defines six "stages of algorithms" for AI (Nowak 2017, 194). The idea is the following: According to Kant, a robot agent should ask itself if the decision he is about to make can become "a universal law for all autonomous agents including myself situated in analogous practical circumstances" (Nowak 2017, 194). Nowak's model is determined by one crucial assumption that can be questioned: just like Kant, her approach considers that moral decisions are the result of rational reflection, deliberation and reasoning. However, Korsgaard – whom Nowak also quotes – shows the opposite. On the basis of Kant's transcendental philosophy, she argues that human beings understood as ends in themselves appear either in an active or in a passive sense. Kant and Nowak concentrate only on the active sense which implies that something is an end in itself if it is a law-maker (capable of rational, and therefore moral choice) (Schmiljun 2018, 54). But human beings are not "merely rational beings" (Korsgaard 2012, 11) and "one person's belief in the morality of their conduct does not necessarily extend to others" (Ulgen 2017, 64). There is a passive dimension, too, that regards us as beings who have certain interests and needs we want to achieve like getting

some food and something to drink. Thus, Nowak's set of algorithms would work for robots, only if all other moral agents acted on the same conditions. Otherwise, Nowak's Kantian machine would be isolated, due to it conflicting with other moral agents who don't constantly act rationally.[3]

In contrast, bottom-up strategies of morality dispense principles and focus on the context sensitive character of morality and situational judgement (Misselhorn 2018, 114). Instead of evaluating the "rightness and wrongness of individual acts", "morality is asserted to be about the character of person" (Abney 2014, 37), Abney notices. This strategy applies especially to virtue-based ethics, asking not 'What should I do?' according to some rules, but it stresses the question 'What should I be' (Gips 2018, 249). It assumes that moral virtues can't be acquired by moral principles. Being a good person is the result of habits, experience and training (Misselhorn 2018, 114). Besides, if I am morally wrong or right depends on the context of the certain situation.

For example, it would be morally wrong if I drive too fast on the street. Another person – let's say a police officer – performing the "same action and in same circumstances" (Abney 2014, 37) can be morally right and virtuous. Maybe the police officer needs to drive fast in order to catch a bank robber. His function in society as someone who defends the law and the situation allow him to drive faster than permitted. In my case, the conditions are different. Virtues can be understood as dispositions to act in a "certain way (…) to know by practical wisdom the right thing to do, in the right way, at the right time" (Abney 2014, 37). Without consulting an abstract universal law or rule, virtue-based ethics ask about changes in habits and character: If I do this, will I still be good person? Bruce McLaren (2003) invented an algorithm based on machine learning called *Truth Teller.* His machine uses data from the judgements of boards of ethical advisors, considering hundreds of cases. *Truth Teller* makes his judgement comparable to doctors or judges. It analyzes a certain case like a doctor takes a look at clinical examples or a judge interprets case histories (Leben 2019, 47).

The challenges appearing at this point are the following: first, virtue-based ethics require adaptive systems that are able to improve their habits and standard settings in regard to new adopted experiences. They must be dynamic and flexible. Second, Misselhorn criticizes rightly that virtue-based systems can contain prejudices that lead to unintended attitudes like discriminating people because of their skin color or race. Third, such algorithms are often non-transparent und difficult to understand for someone standing outside the box.

Apart from these objections, for some disciplines like connectionism, this strategy is still promising (Misselhorn 2018, 115). It shares the idea that moral behavior is more a question of learning and experience than a question of moral principles. Connectionism

3 My objection is probably only half correct. As Ulgen explains, Kant's practical philosophy contains the "understanding and accommodation for the possibility of irrational conduct and wrongdoing" (Ulgen 2017, 64).

is an "approach to neural-networks-based cognitive modeling that encompasses the recent deep learning movement in artificial intelligence" (Kiefer 2019, iv). Misselhorn admits that bottom-up strategies might be a useful instrument to understand the evolution of moral behavior (Misselhorn 2018, 117). Her concern about the question of responsibility towards bottom-up systems seems in my opinion groundless. The question of responsibility is decided on another level – namely in social contexts – in comparison to the question of realization. If my three-years-old daughter has broken a glass on the table, I know that it was her although I probably do not know much about the correct cognitive processes appearing in her brain. The same goes for robots. Once they act autonomously and can be seen as a "robust moral agent", the question of algorithms is secondary when they approach or exceed the "moral status of human beings" with "corresponding rights and responsibilities" (Sullin 2006, 29).

3. Moral Competence in Robot Ethics

In robot ethics, an interdisciplinary discourse of computer scientists, engineers, linguists, psychologists and philosophers have developed computational models (Scheutz, et al. 2017; McLaren 2018; Guarini 2018; Bringsjord *et al*. 2018; Andreae 1987) and several experiments (Scheutz *et al*. 2013) in order to realize a non-biological cognitive system capable of acting morally and consciously.

Thanks to the Multidisciplinary University Research Initiative (MURI), which created the research project *Moral competence in Computational Architectures for Robots,* important groundwork has been done over the last few years (Malle 2014, 2016; Scheutz 2016, 2017; Scheutz & Malle 2014; Scheutz *et al.* 2015). Basically, Malle emphasises that moral competence does not "resolve all ethical concerns over robots in society, but it may be a prerequisite to resolve at least some of them" (Malle 2016, 243). Scheutz adds that "agent designers (need to) to ensure that autonomous artificial agents are equipped with the moral and ethical competence to negotiate human societies in order to prevent the harm they could otherwise cause by being oblivious to ethics and morality" (Scheutz 2017, 57).

Compared to former works (Allen 2011; Moor 2006), Malle and Scheutz refer to moral competence as a result of "elements" (Malle 2016, 245) or "components" (Scheutz 2017, 61). As Malle states, this approach has the advantage that we no "longer need to make tough decisions about whether robots do or do not meet a particular criterion to count as ''fully'' moral agents (Malle 2016, 245; Floridi 2018; Sullin 2006). As such, robots might be individually designed, depending on the specific ''application'' that tells us which ''competences'' must be implemented (Malle 2016, 245). Malle suggests five components for moral competence: Moral decision making and action, moral cognition, moral communication, a system of norms and moral vocabulary (Malle *et al*. 2014, 190; Malle 2016, 245). His concept is based on several discoveries in psychology, neurosciences

and philosophy (Kohlberg 1964; Greene *et al.* 2004; Antaki 1994; McCullough *et al.* 2013). However, the challenge with this approach lies in the problematic definition of moral competence as a result of different components that we can treat as ''separable objects'' (Lind 2016, 51) and that are ''dynamic and adjustable'' (Malle 2016, 245).

In pedagogy and psychology, this argumentative method is not new. We can find similar models that assume moral behavior as an outcome of components containing affect (moral orientation), cognition (moral cognition) and behaviour. ''The processing structure of the disposition consists of a characteristic set of cognition, affects, and behavioural strategies in an organization of interrelations that guides and constrains their activation.'' (Mischel & Shoda 1995, 257) Lind objects that these components sometimes are referred to as separate ''domains'', ''processes'', ''parts'', ''sets'' or ''sub-systems'' of behavior (Lind 2016, 51). As Greene, for example, puts it: ''We are explaining moral thinking in terms of its more basic cognitive components" (Greene 2015, 40). Further, Greene and his associates speak of the cognitive and emotional processes underlying moral behavior as ''competing subsystems'' (Greene 2004, 389). Moreover, Jim Rest provides a definition of moral behaviour, consisting of four components: moral sensitivity, moral judgement, moral motivation, moral character (Rest *et al.* 1999, 101).

Though Malle and Scheutz explicitly refer to moral competence, not to moral behaviour, it can be noticed that they use a similar terminology which might be misplaced here. It suggests that moral decision making and action, moral cognition, moral communication etc. are some kind of objects that we can randomly change in a robot or neglect. ''Robots may be designed to have some competences but not others'' (Malle 2016, 245).

4. The Dual Aspect Dual Layer Theory of the Moral Self

In contrast to this component theory, I will propose another perspective of moral competence developed by Georg Lind, which – as I will demonstrate – meets several additional conditions and goals that also Scheutz and Malle would agree with. According to Lind, as mentioned above, moral competence is the ability to ''resolve problems and conflicts on the basis of inner moral principles through deliberation and discussion instead of violence and deceit" (Lind 2016, 13). This applies to Scheutz's remark, "robots will need mechanisms analogous to humans to deal with the situational openness and unpredictability of human societies: they need to be explicit ethical agents, able to represent, learn and reason with norms and values in much the same way humans do" (Scheutz 2017, 62).

Lind's definition of moral competence is based on the former works of Kohlberg, Piaget and others (Lind 2016, 45). They concentrate on the "underlying cognitive structure" of moral behavior (Lind 2016, 45) and believe that "moral–judgmental structures in a human mind" (Nowak 2016, 322) can be "constructed through the individual´s interaction

with his or her social environment" (Lind 1985, 43). Explicitly, Lind sees his model in the tradition of Habermas. According to Lind and Habermas, "competence by itself cannot be shown to exist except in its concrete manifestation, that is, through phenomena of performance" (Lind *et al.* 1985, 25). Given this, Lind develops his *Dual Aspect Theory* that understands moral competence only as one aspect **(cognitive aspect)** of the moral self. The other aspect refers to **moral orientation (affective aspect).** Furthermore, both aspects can be distinguished **into two layers**. I will exclude this final point in the subsequent discussion, as it would raise a new philosophical question of whether robots are able to be conscious (Bringsjord *et al.* 2018; Andreae 1987).

To start with the latter aspect **(affective aspect)**, Lind understands moral orientation as the "preferred moral principle which a person expresses in her or his patterns of responses" (Lind 2016, 184). According to Kohlberg and Lind, we can define six types and three categories of moral orientations (Lind 2016, 53 and Kohlberg 1964, 400). These categories include "pre-conventional", "conventional" and "principled-conventional" orientations. All six types are linked with ethical reasons of moral action, such as "Avoid physical damage and injury to oneself" (which is pre-conventional because it isn't normative in its nature; it is factual or experiential as the feeling of pain one wants to avoid), or normative (conventional) reasons such as "respect the laws and the order of society and contribute to its maintenance", or post-conventional reasons, i.e., more general than any kind of society-, culture-, confession- etc. related rules, norms and conventions. "Hold up universal principles of justice, reason and logic" belongs to the post-conventional order. All six types of moral orientation are universal for and part of our cognitive structure; "they are our common heritage" (Lind 2016, 16).[4]

This aspect resembles the previously mentioned top-down strategy, "integrating rules and principles into the architecture" (Misselhorn 2018, 96; Scheutz 2017, 60) of a robot. It is also close to Malle's "system of norms" that a "community adopts to regulate individual community members' behaviors and thus bring them in line with community interests" (Malle 2016, 246). Like Lind and Kohlberg, Malle speaks too of a "hierarchy of norms" (Malle 2016, 246).

The second aspect **(cognitive aspect)**, moral competence, on the other hand, demonstrates "proficiency and virtuosity in making demanding and new context-related decisions" (Nowak 2016, 324). Moral competence is an "ability to apply a certain moral orientation in a consistent and differentiated manner in varying social situations" (Prehn *et al.* 2007, 44). It involves "ethical judgment and reasoning" (Lind 2016, 57) that will finally conclude with "concrete moral judgment, or decision" (Nowak 2016, 329). In addition, Lind is convinced that human beings are able to learn to "discursively express, improve (to make them just), and to justify their judgments to significant extent" (Nowak 2016, 329). Malle seems to have the same in mind when he states that "moral competence

4 Lind gives a more detailed explanation of all three categories in his dissertation (compare Lind 1985, 61).

is an aptitude, a qualification, a dispositional capacity to deal adequately with certain tasks" (Malle 2016, 245). Even his two elements of "moral decision making" and "moral cognition and affect" reveal parallels to Lind's cognitive aspect. For example, Malle understands moral cognition as a process of "perception and judgment" (Malle 2016, 248) to detect and evaluate norm-violating events and actions and to respond to them. Further, Malle explains that "moral decision making" includes "affective states and personality dispositions, automatic imitation and group pressure, heuristics and reasoned choice" (Malle 2016, 249). To Lind moral competence is a skill that allows someone to become aware of problems and to solve them through thinking. The greater a problem, the better this ability must be developed (Lind 2016, 45). Besides this, it is possible that people react to a dilemma merely on a pre-conventional stage, or as Malle calls it "nonmoral behavior" (Malle 2016, 249), for example, when we are asked if we want to have wine or soda to drink at dinner (Lind 2016, 45). Lind points out that it is ok "when we consider which beverage agrees better with us" (Type 2: Acquire benefits and rewards). Discussing this issue on a higher level seems relatively unnecessary.

As illustrated, situated judgments and the adaptive behavior of cognitive systems are properties found in bottom-up strategies (Misselhorn 2018, 114). Moral competence would include that robots can handle new context situations and apply their decisions and judgements based on learned principles, rules or norms. Scheutz notes that a model of moral competence requires a "sufficient computational understanding (how) humans learn, represent, and reason with moral norms, and how they detect, violate norms themselves, and respond to norm violations from others" (Scheutz 2017, 61). This aspect would require Malle's two other elements, namely "moral vocabulary" and "moral communication", as "cognitive tools" (Malle 2016, 251).

Lind addresses his *Dual Aspect Dual Layer Theory* first of all to human beings which could make it difficult to transfer his approach to robots, including its computational realization. However, it contributes a new framework to moral competence, which, on the one hand, avoids a component theory, and on the other hand, could bridge the gap between top-down and bottom-up strategies in robotics. Moreover, Lind clarifies that moral competence has to be understood as one aspect of morality: It is not a component of behaviour. I already tried to outline the many parallels between the two different concepts of Malle and Lind. I think that these can be merged, although many questions appear, such as: Do we need all six stages of moral orientation for robots? Maybe it is more sensible to program only the "conventional" and "principled-conventional" stages? How can we translate all moral orientation into a formalized language, into algorithms? In which robot application is an integration of moral competence sensible? Obviously, it would be unnecessary and inefficient to implement moral competence into an autonomous vacuum cleaner.

Scheutz is right that we stand at the very beginning of our research to develop a humanlike morality in robots although there are already a few current experiments.

(Scheutz 2017, 61) Moreover, Scheutz shows his proximity to Lind: "The goal is to develop explicit representations of social and moral norms, as well as inference, decision-making, and action-execution algorithms, that will allow robots (1) to detect morally charged situations, (2) reason through them based on their ethical rules, norms, obligations and permissions, and (3) find the best action that meets their obligations while minimizing harm to humans" (Scheutz 2017, 63). I believe Lind would agree with this description of a morally competent agent too.

5. Conclusion

McLaren points out that "imbuing a computer with the ability to reason about ethical problems and dilemmas is as difficult a task as there is for Artificial Intelligence (AI) scientists and engineers" (McLaren 2018, 297). As shown, top-down and bottom-up strategies are promising but don't provide sufficient solutions for robots to cope with other moral (human) beings in society. In comparison, the concept of moral competence appears to be a sensible alternative that we should foster in robots. Surely, we need to consider in which application domains an integration of moral competence is valuable. I believe that especially interactions between robots and children in educational environments or interactions between robots and elderly people in hospitals could be prime candidates. For this reason, I agree with Scheutz about the importance of this project, although I wouldn't use the same negative and alarming words: "failing to develop appropriate cognitive architectures for autonomous agents that are sensitive to human ethical and moral concerns as well as our social emotional needs could turn utopia into dystopia" (Scheutz 2017, 63). A human moral competence can be regarded as a chance to "minimize human harm" (Scheutz 2016, 518). With respect to Lind's discoveries, we have a framework of how to implement a humanlike moral competence into robots. One thing is sure, autonomous robots will be faced with similar conflicts in society to us. If we don't want them to be a risk and danger for future generations, robots must be able to resolve conflicts on the basis of principles through deliberation and thinking, instead of violence, force and deceit.

References

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

Allen C. 2011. "The Future of Moral Machines." The New York Times: Opinionator. Retrieved December 29, 2014, from http://opinionator.blogs.nytimes.com/2011/12/25/the-future-of-moral-machines/.

Antaki C. 1994. *Explaining and Arguing: The Social Organization of Accounts*. London: Sage.

Anderson M. & Anderson S. L. 2018. "General Introduction." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 1-4). Cambridge: University Press.

Andreae John H. 1987. "Design of Conscious Robots." *Metascience* 5:41-54.

Bringsjord S., Taylor J., Van Heuveln B., Arkoudas K., Clark M., & Wojtowicz R. 2018. "Piagetian Roboethics via Category Theory: Moving beyond Mere Formal Operations to Engineer Robots Whose Decisions Are Guaranteed to Be Ethically Correct." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 361-374). Cambridge: University Press.

Floridi L. 2018. "On the Morality of Artificial Agents." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 184-212). Cambridge: University Press.

Gips J. 2018. "Towards the Ethical Robot." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 244-253). Cambridge: University Press.

Guarini M. 2018. "Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism-Generalism Debate." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 316-334). Cambridge: University Press.

Greene J. D., Nystrom L. E., Engell A. D., Darley J. M., & Cohen J. D. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44:389-400

Greene, J. D. 2015. "The Rise of Moral Cognition." *Cognition* 135:39-42.

Hall J. Storrs 2018. "Ethics for Machines." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 28-44). Cambridge: University Press.

Kant I. 1993/1785[1]. *Grounding for the Metaphysics of Morals*. Trans. by J. W. Ellington (3rd ed.). Indianapolis – Cambridge: Hackett.

Kiefer A. B. 2019. *A Defense of Pure Connectionism.* Diss. The Graduate Center, City University of New York. DOI:10.13140/RG.2.2.18476.51842.

Kohlberg L. 1964. "Development of Moral Character and Moral Ideology." In: M. L. Hofmann & L. W. Hoffmann (Eds.), *Review of Child Development Research* (pp. 383-432). New York: Russell Sage Foundation.

Korsgaard C. M. 2012. "A Kantian Case for Animal Rights." In: M. Michel, D. Kühne, & J. Hänni (Eds.), *Animal Laws – Tier und Recht. Developments and Perspectives in the 21st Century* (pp. 3-27). Zürich – St. Gallen: DIKE.

Leben D. 2019. *Ethics for Robots. How do Design a Moral Algorithm*. London – New York: Routledge.

Lin P. 2014. "Introduction into Robot Ethics." In: P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics* (pp. 3-16). Cambridge: MIT University Press.

Lind G. & Wakenhut R. H. 1985. "Testing for Moral Judgment Competence." In: G. Lind, H. A. Hartmann, & R. Wakenhut (Eds.), *Moral Judgments and Social Education* (pp. 79-105). New Brunswick –London: Transaction Publishers.

Lind, G. 1985. *Inhalt und Struktur des moralischen Urteilens. Theoretische, methodologische und empirische Untersuchungen zur Urteils- und Demkratiekompetenz bei Studierenden.* Diss. Konstanz: Universitätsdruck.

Lind G. 2016. *How To Teach Morality. Promoting Deliberation and Discussion, Reducing Violence and Deceit*. Berlin: Logos Verlag.

Malle B. F. 2014. "Moral Competence in Robots?." In: J. Seibt, R. Hakli, & M. Nørskov (Eds.), *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy* (pp. 189-198). Amsterdam, Netherlands: IOS Press. DOI:10.3233/978-1-61499-480-0-189.

Malle B. F. 2016. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18:243-256. DOI:10.1007/s10676-015-9367-8.

McLaren B. M. 2018. "Computational Models of Ethical Reasoning. Challenges, Initial Steps, and Future Directions." In: M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 297-315). Cambridge: University Press.

McCullough M. E., Kurzban R., & Tabak B. A. 2013. "Putting Revenge and Forgiveness in an Evolutionary Context." *Behavioral and Brain Sciences* 36:41-58. DOI:10.1017/S0140525X12001513.

Millar J. 2017. "Ethics Settings for Autonomous Vehicles." In: P. Lin, R. Jenkis, K. Abney (Eds.), *Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence* (pp. 20-34). Oxford: Oxford University Press.

Mischel W. & Shoda Y. 1995. "A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure." *Psychological Review* 102(2):246-268. DOI:10.1037/0033-295X.102.2.246.

Misselhorn C. 2018. *Grundfragen der Maschinenethik*. Stuttgart: Reclam.

Moor J. H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 2:18-21. DOI:10.1109/MIS. 2006.80.

Nowak E. 2016. "What Is Moral Competence and Why Promote It?." *Ethics in Progress* 7(1):322-333. DOI:10.14746/eip.2016.1.18

Nowak, E. 2017. "Can Human and Artificial Agents Share an Autonomy, Categorical Imperative-Based Ethics and 'Moral' Selfhood?" *Filozofia Publiczna I Edukacja Demokratyczna* 6(2):169-208. DOI:10.14746/fped.2017.6.2.20

Prehn K., Wartenburger I., Meriau K., Scheibe Ch., Goodenough O. R., Villringer A., van der Meer E., & Heekeren H. R. 2008. "Individual Differences in Moral Judgment Competence Influence Neural Correlates of Socio-normative Judgments." *Social Cognitive and Affective Neuroscience* 3:33-46.

Rest J. R., Narváez D., Bebeau M. J., & Thoma S. J. 1999. *Postconventional Moral Thinking. A Neo-Kohlbergian Approach.* Mahwah, NJ: Erlbaum.

Scheutz M., Briggs G., Cantrell R., Krause E., Williams T., & Veale R. 2013. "Novel Mechanisms for Natural Human-Robot Interactions in the DIARC Architecture." *Intelligent Robotic Systems: Papers from the AAAI* 2013 Workshop: 66-72.

Scheutz M. & Malle B. F. 2014. ''Think and Do the Right Thing': A Plea for Morally Competent Autonomous Robots." Presented at the 2014 IEEE Ethics Conference, Chicago, IL. DOI:10.1109/ETHICS.2014.6893457

Scheutz M., Malle B. F., & Briggs G. 2015. „Towards Morally Sensitive Action Selection for Autonomous Social Robots." Presented at the 2015 IEEE Ethics conference, Kobe, Japan. DOI:10.1109/ROMAN.2015.7333661.

Scheutz M. 2016. ''The Need for Moral Competency in Autonomous Agent Architectures." In: V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 517-527). Heidelberg: Springer.

Scheutz M. 2017 (Winter). "The Case for Explicit Ethical Agents." *AI Magazine* 38(4):57-64. DOI:10.1609/aimag.v38i4.2746

Scheutz M., Baral C., & Lumpkin B. 2017. "A High Level Language for Human Robot Interaction." *Advances in Cognitive Systems* 5:1-16.

Schmiljun A. 2017. "Robot Morality: Bertram F. Malle's Concept of Moral Competence." *Ethics in Progress* 8 (2):69-79. DOI:10.14746/eip.2017.2.6

Schmiljun, A. 2018. "Why Can't We Regard Robots As People?." *Ethics in Progress* 9(1):44-61. DOI:10.14746/eip.2018.1.3

Steć M. 2017. "Is The Stimulation of Moral Competence with KMDD® Well-suited for Our Brain? A Perspective From Neuroethics." *Ethics in Progress* 8(2):44-58. DOI:10.14746/eip.2017.2.4 .

Sullin J. P. 2006. "When Is a Robot a Moral Agent?." *International Review of Information Ethics* 6(12):23-30.

Ulgen O. 2017. "Kantian Ethics in the Age of Artificial Intelligence and Robotics." *QIL* 43:59-83.

Veruggio G., Solis J., & Van der Loos, M. 2011. "Roboethics: Ethics Applied to Robotics." *IEEE Robotics & Automation Magazine* 18(1): 22-23.

Wallach W. 2010. "Robot Minds and Human Ethics: The Need for a Comprehensive." *Ethics and Information Technology* 12(3):243-250. DOI:10.1007/s10676-010-9232-8.

André Schmiljun (Berlin)

Moral Competence and Moral Orientation in Robots

Abstract: Two major strategies (the top-down and bottom-up strategies) are currently discussed in robot ethics for moral integration. I will argue that both strategies are not sufficient. Instead, I agree with Bertram F. Malle and Matthias Scheutz that robots need to be equipped with moral competence if we don't want them to be a potential risk in society, causing harm, social problems or conflicts. However, I claim that we should not define moral competence merely as a result of different "elements" or "components" we can randomly change. My suggestion is to follow Georg Lind's *dual aspect dual layer theory of moral self* that provides a broader perspective and another vocabulary for the discussion in robot ethics. According to Lind, moral competence is only one aspect of moral behavior that we cannot separate from its second aspect: moral orientation. As a result, the thesis of this paper is that integrating morality into robots has to include moral orientation and moral competence.

Keywords: moral competence; moral orientation; Georg Lind; robot ethics; Dual Aspect Dual Layer Theory of Moral Self.

Ethics in Progress (ISSN 2084-9257). Vol. 10 (2019). No. 2, Art. #9, pp. 98-111.

DOI:10.14746/eip.2019.2.9