Ethics in Progress (ISSN 2084-9257). Vol. 8 (2017). No. 2,
Art. #6, pp. 69-79. Doi: 10.14746/eip.2017.2.6

Robot Morality: Bertram F. Malle's Concept of Moral Competence

André Schmiljun (Humboldt University in Berlin, andre_schmiljun@icloud.com)

1. Introduction

Nowadays, robotics is a rapidly increasing industry producing new developments year after year. In 2007, Bill Gates observed: "The emergence of the robotics industry is developing in the same way that the computer business did 30 years ago" (Lin 2014, 1). Those robots already clean our houses, mow our lawns, hunt terrorists or transport heavy loads weighing up to 1.000 kg. Over the course of the last few years, robot usage in society has expanded enormously and they now carry out a remarkable number of tasks for us. In every industrial sector, it is likely that there are at least a handful of jobs for human workers that will sooner or later be replaced by robots or autonomous solutions (Lin 2014, 1).

Currently, robots are mostly tasked with duties that are seen as non-value adding, exceptionally dull or even dangerous. They are considered as means to support and substitute human workers where those are handicapped or limited. For instance, automobile factory robots execute the same, repetitive assemblies again and again 24 hours a day without any break, with precision and perfection; military unmanned aerial vehicles surveil and control from the skies for far more hours than a human pilot can endure at a time. In logistics, robots carry packages, palettes and barrels through difficult, complex areas with reliability and precision and collaborate with human workers. Without any fear of danger or risk, they also explore volcanoes, travel to Mars, secure contaminated sites and defuse bombs. It is not surprising that, four years ago, Linda Johansson noted in her Doctoral thesis: "We read about them (robots) in the newspapers almost every day. (…) When we make a phone call to a company and get to talk to a computer, it seems like the world is becoming more and more automated" (Johansson 2013, 1).

Linked with the rise of robotics is the question of morality. As robots become more autonomous[1] (Johansson 2013, 1), it perhaps becomes plausible to assign responsibility to the robot itself rather than its creator, especially if it is able to meet with most of the features that typically define personhood. A popular scenario frequently quoted by moralists illustrates the dilemma: If a human driver causes an accident, the driver has to face the consequences of his carelessness. He is responsible for what he has done. But if it was the car driving autonomously without any human interference, the situation is different. And what if an accident is unavoidable and the car has to decide whether to save the passengers in it or uninvolved people on the street (Bendel 2013)?

The philosopher Bertram F. Malle therefore calls for a debate of *moral competence in robots.* As he puts it: Any robot that collaborates with, looks after or helps humans is a social robot that must have moral competence. He outlines moral competence as a functional system of five cognitive abilities that seems to put machines on the same level as humans. Starting with a short introduction to robot morality, I will analyse Malle's five components of moral competence and will discuss in how far his approach tangles with common ideas of personhood.

2. Robot Morality

Robot morality, or, as it is correctly called, *robot ethics* is a very young discipline. Many authors yet criticize that it does not have a specific object of research due to ethics normally addressing animated matter such as animals or humans (Loh 2017, 22). However, most are willing to admit that ecosystems, cars, houses, smartphones and a range of various other entities have a value.

The term "robot" originally refers to the Czech word "robota", meaning work and compulsory labour and was introduced by the artist Josef Čapek in 1920. In his play "Rossum's Universal Robots" (1921), his brother Karel Čapek spoke about "labori" for humanoid equipment serving humans to ease their work. Literature gives many definitions of what robots are. But generally, it can be agreed that typically, a robot uses sensors to detect aspects of an external world, software to reason about it, and actuators to interact with it. We can thus define robots as a branch of engineering that deals with autonomous machines (Abney & Veruggio 2014, 349-50).

[1] The term „autonomous" is not defined by exact definition. Being autonomous is linked with the basic idea to have the ability to be off on one's own, making decisions of one's own, without the influence of someone else.

Generally, Robot ethics discuss the questions if robots do have a moral value and in which way they can be seen as a moral subject. It discusses the question of which components are essential for *moral agency* and what moral code we want to programme into them somehow (Abney & Veruggio 2014, 347). Say it becomes necessary to write software for an autonomous robot collaborating with us in a certain context, then we must decide which moral values ought to be followed by it and which ones it does not need. This decision will enable humans to judge whether the robot acts morally, in that it either obeys its programmed moral code and does what it ought to do, or acts immorally, doing something it is not supposed to do, be it because of an electromechanical glitch, or a bug in its software. Finally, Robotic ethics ask how we as human beings should treat robots and what it means if we act "unfair" towards them. In which industrial sector or society area do we want artificial support: In medicine and healthcare, in military, in research and education or in waste management (Johansson 2013, 67-82)?

Robot ethics can be approached by examining two categories: robots as *moral patients* and robots as *moral agents* (Loh 2017, 22). The first category considers robots as passive holders of moral rights, functioning as objects of moral responsibility of moral agents. Moral attitudes such as concern, respect, or care can be directed at moral patients and moral agents can have moral responsibilities towards them. In this understanding of moral actors, all moral agents are also moral patients, but moral patients need not to be moral agents. It is only the moral agent who is an active holder of moral obligations and responsibilities (Winston 2008). Analysing robots as part of the first category (as moral patients), robot ethics mainly asks about the correct human behaviour when it comes to their application. In this case, artificial systems are predominantly understood as tools or technical supplements for humans. Analysing robots as part of the second category (as moral agents), scientist refer to them as subjects having the individual's ability to make moral judgements based of some simple notion of what is right or wrong. Here, Malle criticizes that many scholars mix up moral agency with moral competence (Malle 2014, 189). In his opinion, moral competence goes further, as we will see in the next chapter.

3. Moral Competence

Thanks to Lawrence Kohlberg and Georg Lind, a lot of research has been conducted on the term "moral competence" over the course of the last few years. Kohlberg defines moral competence as the cognitive ability to make judgments and decisions that are based on internal moral principles, and to act in accordance with such judgments (Kohlberg 1964, 425). Lind on the other

hand states: “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).

Taking one step back, the word competence originates from the Latin word *competentia*, which in post-classical Latin was combined with the meaning of “meeting together” or “agreement”, but it also stems from *competere*, which means to compete or rival (Malle 2014, 189). We see that the word competence is set in between contrasts: namely competition and cooperation. In common sense, competence is considered an aptitude, a qualification, a dispositional capacity to deal adequately with certain tasks (Malle 2014, 189).

According to Malle, moral competence is a set of five components: (i) a system of norms, (ii) a moral vocabulary, (iii) moral cognition and affect, (iv) moral decision making and action, (v) moral communication. Following his interpretation, we can speak of moral competence only if all of these criteria are fulfilled. Hence, Malle’s main argument resembles general theories of functionalists (Beckermann 2001, 143-45) who usually put their arguments in the same logical way, saying: If a subject or system is in a certain state (a), and something (internal or external) is added to the subject or system, state (a) will change into state (b). A popular example for this is a vending machine. If the vending machine is in a certain state (a) and someone inserts money, the vending machine ejects a can. Transferring it to our case of moral competence: If a robot is added all five essentials, it will be in a state that would allow us to call it morally competent. Like many functionalists, Malle believes we can treat cognitive abilities as a phenomenon that is not only reserved for humans, but is something we can implement into a machine as complex software, once we know how it works.

3.1 A system of norms

The philosopher Immanuel Kant defined morality as compliance with universally valid moral principles instead of a simple list of prohibitions and commandments for behaviour. To him, moral principles were maxims of actions that we would wish to be universally valid and applied (Lind 2016, 61). We act morally if our behaviour coincides with our principles. Malle seems to agree with Kant’s point of view that morality is necessary to regulate human social living. He believes that human communities perform this regulation by motivating and deterring certain behaviours through the imposition of norms (Malle 2014, 190).

Thus, the first essential characteristic of moral competence is a system of norms, although there are still many unanswered questions in human psychology about those. For instance, we do not know exactly how norms are acquired or represented in the human mind, what properties they have that

allow them to be so context-sensitive and mutually adjusting. Regarding the development of norms, it is known that children are able to express concrete moral judgements (Wright & Bartsch 2008, 56-85): "What you have done was wrong! This is not nice." However, children also easily induce more general rules from concrete instances, such as "bombs hurt people" (Malle 2014, 190).

Malle points out that norms function like goal concepts somewhat, which can typically be found in robot architectures. Depending on the robot, on a very simple level, goals provide the robot with parameters, for instance where to go, what to carry and generally what to do. Malle further suggests that moral norms require a more complex goal definition. A norm needs to have the quality of representations and value as he follows Jon Elster's opinion that "social norms provide an important kind of motivation for action that is irreducible to rationality or indeed any other form of optimizing mechanism" (Elster 1989, 15). But if we understand how representations are constituted in the human mind and how contextual activities allow us to detect norm violations, Malle is convinced that it is possible to build computed norm systems.

Malle puts this very simply. Neurosciences, biology and philosophy yet struggle to find a convincing concept of the mind and are far from answering fundamental epistemic questions such as if representations are really internal or external.[2] But this question is important before we start with technical development of norm systems in robots. And, apart from that, we must ask: What set of norms do we implement? Do we use all existing human norms on the planet, or do we evaluate country by country?

3.2 Moral vocabulary

The second essential ingredient needed for a human or robot to be morally competent is a steady moral vocabulary. Malle's premise is that a norm system demands language for learning it, using it, and negotiating it. In terms of humans, he might be right. But why would we have to implement an entire human language in a machine? Why not use simple binary code? Why should we implement millions of words with millions of connotations into a robot when a binary code is more precise? This huge expanse is only comprehensible if we intend to set up communication based on human abilities.

Malle introduces three categories of moral vocabulary: **Vocabulary of norms and their properties** ("fair," "virtuous," "reciprocity," "honesty," "obligation," "prohibited," "ought to," etc.); **Vocabulary of norm violations** ("wrong," "culpable," "reckless," "thief," but also "intentional," "knowingly," etc.); **Vocabulary of responses to violations** ("blame," "reprimand," "excuse," "forgiveness," etc.). In each domain, there are numerous distinctions and differentiations.

[2] I am referring here mainly to the philosophical debates labelled internalism and externalism, as well the whole philosophical school of "New Realism" around Markus Gabriel, which is trying to answer similar questions (Searle 1998; Roth 2007).

Let us assume Malle was right and we were able to create a robot with moral German, English or Polish vocabulary, then how do we guarantee that it uses words and sentences in the same way we do? Especially since moral vocabulary is highly influenced by current culture, society and history?

3.3 Moral cognition and affect

Moral vocabulary on one hand and a moral norm system on the other hand are not enough to explain why we call an incident bad or good and say that anyone deserves blame or praise. So what psychological processes are involved in detecting and responding to norm violations? Malle distinguishes between two types of moral judgements: *events* (outcomes, behaviours) and *agents.* He is correct in that the key difference between the two forms is mainly the amount of information processing that normally underlies each judgment (Malle 2014, 192). An event judgement merely demands that we register that a norm has been violated. If an agent has done something wrong, we usually take the agent's specific causal involvement, intentionality, and mental states into account.

Nevertheless, registering that an event violated a norm is not as simple as it seems. For instance, if a robot sees a police officer killing a dangerous criminal or terrorist, it must be capable of distinguishing that this action was necessary even though it violates the norm to never harm anyone. This situation becomes more complex if a robot has to consider intentions and reasons of the agent. The robot therefore has to understand that many human actions are based on reasons. Unfortunately, Malle puts only a small spotlight on emotions in moral judgement. It was the remarkable work of Daniel Kahneman who demonstrated in many experiments that most human judgment works intuitively and emotionally (Kahneman 2011). Malle briefly discusses the role of affects, but concludes that it is not an important factor for the creation of a morally acting robot (Malle 2014, 193).

3.4 Decision making and action

A fourth element required for moral competence is decision making. Malle mainly limits the debate to two psychological terms: *empathy and self-regulation*. He says that an action becomes moral by the involvement of socially shared norms and individual goals. For instance, it can be the individual goal of an autonomous bus to save its passengers. But if the bus is suddenly involved in an unavoidable accident, it must decide whom it should save: the people in the car or the people on the street. The bus has to make a moral decision depending on what is right or wrong.

Malle is misleading when he thinks he can avoid this problem in designing a robot without any self-regulation and awareness of community benefits. Of course, it depends on what is understood by the word "self-

regulation". According to Malle, self-regulation is all about being self-interested and cold, ignoring other's needs or building trust (Malle 2014, 194). This is, I think, a false and very human idea of robots. Right now, robots already calculate their position, battery status, routes and duties and coordinate all data with other robots. What makes Malle so sure that they are unable to solve problems in cooperation?

We see that Malle's position is contradictory. On one hand he believes that robots can act reasonably, on the other hand he is concerned that robots can't handle self-regulation (following its own goals) and community values at the same time.

3.5 Moral communication

Finally, moral competence is a matter of communication. If we want to regulate other people's behaviour, we need to express what we wish to regulate. The same is true for robots. If someone has made a mistake and violated a norm, we blame them for their decision. Blame in this case functions as a social act to inform, correct and provide an opportunity to learn (Malle, Guglielmo & Monroe 2014, 147-186). I agree with Malle in that robots can change and learn, and that, if they can make decisions by themselves, they become be appropriate targets of blame. In this case a robot is in the same position as any other agent. It can come up with and express moral judgements and can therefore also be accused.

Furthermore, moral competence requires the ability to explain immoral behaviour (typically one's own, but also sometimes others'). Thus, an essential question is how robots are to access their own intentional behaviour and reasoning. Will they be able to know their desires and beliefs in light of which and on the ground of which they decided to act? And if robots become autonomous on such a high level of social behaviour, is it likely that robots might not always truthfully report their internal reasons of their actions? After all, if they have to follow a superior norm or value, they could decide to hide their true intentions and lie to us.

4. Conclusion: Are Machines People?

The last chapter already illustrates the core problem of Malle's whole futuristic theory. Despite all my objections and doubts, his idea of moral competence raises the question if machines can someday be understood as people. To make this point clear, I will shortly summarize the key characteristics of personhood, before I explain my position. Generally, personhood or a person is characterized by his or her individual properties. The word originates from the Latin word *persona*, meaning singularity, uniqueness and individuality.

The philosophical literature on this subject is very extensive. Dieter Sturma for example claims that, as the term has many different meanings, it should be difficult to define it with absolute precision in one exact definition. He points out that a person is someone who lives their life self-determinedly, making moral decisions and following individual plans, ideas and beliefs. (Sturma 1997, 348). On the other hand, Harry G. Frankfurt states that a person is a special entity whose existence is more profound than their biological happenstance. A person has the capacity to properly identify with their desires and has a will. A person can reflect on their inner wishes, reasons and motivations (Frankfurt 1971, 6). Finally, the constitution of personhood depends on community and the presence of others. It was the German philosopher F. W. J. Schelling who pointed out that a person demands the presence of another person. „Und so ist es auch das Ich, welches als selbst Persönliches Persönlichkeit verlangt, eine Person fordert (…) ein Herz, das ihm gleich sey" (Sturma 2015, 67).[3] In the presence of other people, we become aware of our own individual qualities, behaviour and self-interests and learn to coordinate and communicate them in community.

Unsurprisingly, we see that Malle's concept of moral competence matches with a majority of characteristics usually attributed to a person. This means that, if we follow Malle's theory, we have to think about its consequences, too. If a machine acted autonomously and was able to make its own decisions, had access to its reasons and intentions, was able to blame and correct its environment, why shouldn't we treat a machine as a person? Malle doesn't give a proper answer here. He rather concentrates on the issue of how we should integrate robots in society (Malle 2015, 19). But it is not enough to consider only the ethical implications at this point. We have to think about *robot personality or personhood*, too. For instance, if machines were on the same moral level as humans, would they have the same rights and obligations like us? Or should we treat machines differently? Malle answers this question somehow contradictory. On one hand, he can't deny that robots should have moral standing. On the other hand, he explicates that their rights maybe limited, however, and vary as a function of their value and specific role in society. If a robot met all five elements of Malle's moral competence, I believe we would have no other choice but to accept its personality. Nevertheless, this raises new philosophical questions: Will human personality and robot personality be the same? How will robots experience their personality? What does it mean for a robot's personality that it doesn't age? And if a robot died, would it have the same meaning as the death of a human being? Rob Sparrow puts it this way: "Machines will be people when we can't let them die without

[3] Unfortunately, the original source of this Schelling quotation – written in the essay „Dartellung der reinrationalen Philosophie" - is not yet published.

facing the same moral dilemma that we would when thinking about letting a human being die" (Sparrow 2014, 307).

To sum, although Malle's approach is very unique and innovative, it is insufficient. The nature of his main argument, treating moral competence like software we can easily implement into a robot, is putting a complex subject too simply. Therefore, every component of his concept is linked with new questions and problems and in the end confronts us with the challenging question of personhood in machines. I believe that Malle's thoughts aim in the right direction, but his ideas require more discussion and debate.

References

Abney K. & Veruggio G. 2014. "Roboethics: The Applied Ethics for a new Science," in P. Lin (Ed.), *Robot Ethics (Intelligent Robotics and Autonomous Agents).* Cambridge: MIT University Press (347-364).

Beckermann A. 2001. *Analytische Einführung in die Philosophie des Geistes*. Berlin: Walter de Gruyter.

Bendel O. 2013. "Wie viel Moral muss eine Maschine haben?" Liewo. Retrieved from http://blog.zdf.de/hyperland, on August 12, 2017).

Elster J. 1989. *The Cement of Society: A Study of Social Order*. New York, NY: Cambridge University Press.

Frankfurt H. G. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68 (1): 5-20.

Johansson L. 2013. *Autonomous Systems in Society and War: Philosophical Inquiries*. Stockholm: KTH Royal Institute of Technology.

Johansson L. 2013. "Robots and the Ethics of Care." *International Journal of Technology* 4 (1):67-82.

Kahneman D. 2011. *Schnelles Denken, Langsames Denken*. München: Pantheon.

Kohlberg L. 1964. "Development of Moral Character and Moral Ideology," in M. L. Hoffman & L. W. Hoffman (Eds.), *Review of Child Development Research*, Vol. I. New York: Russel Sage Foundation (381-431).

Lin P. 2014. *Robot Ethics (Intelligent Robotics and Autonomous Agents).* Cambridge: MIT University Press.

Lind G. 2016. *How to Teach Morality?* Berlin: Logos Verlag.

Loh J. 2017. "Roboterethik. Über eine noch junge Bereichsethik." *Information Philosophie* 1: 20-33.

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*. Series: Frontiers in Artificial Intelligence and Applications. Ios Pr Inc 273 (189-198).

Malle B. F. 2015. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18 (4):243-256. doi: 10.1007/s10676-015-9367-8.

Malle B. F., Guglielmo S., & Monroe A. E. 2014. "Moral, Cognitive, and Social: A Theory of Blame." *Psychological Inquiry* 25 (1):147-186.

Roth, G. 2007. *Persönlichkeit, Entscheidung und Verhalten. Warum es so schwierig ist, sich und andere zu ändern*. Stuttgart: Klett-Cotta.

Searle J. R. 1998. *Geist, Sprache und Gesellschaft*. Frankfurt/Main: Suhrkamp.

Sparrow R. 2014. "Can Machines Be People? Reflections on the Turing Triage Test," in P. Lin (Ed.), *Robot Ethics...* (301-316).

Sturma D. 1997. *Philosophie der Person. Die Selbstverhältnisse von Subjektivität und Moralität*. Paderborn – München – Wien – Zürich: Mentis.

Sturma D. 2015. "Person sucht Person," in T. Buchheim & F. Hermanni (Hrsg.), *Alle Persönlichkeit ruht auf einem dunkeln Grunde*. München: De Gruyter.

Winston M. 2008. "Moral Patients." Retrieved from http://ethicsofglobalresponsibility.blogspot.de/2008/02/moral-patients.html on August 13, 2017 (no pages).

Wright J. C. & Bartsch K. 2008. "Portraits of Early Moral Sensibility in Two Children's Everyday Conversation." *Merrill-Palmer Quarterly* 54 (1): Article 4.

André Schmiljun (Humboldt University in Berlin)

Robot Morality: Bertram F. Malle's Concept of Moral Competence

Abstract: Bertram F. Malle is one of the first scientists, combining robotics with moral competence. His theory outlines that moral competence can be understood as a system of five components including moral norms, a moral vocabulary, moral cognition, moral decision making and moral communication. Giving a brief (1) introduction of robot morality, the essay analyses Malle's concept of moral competence (2) and discusses its consequences (3) for the future of robot science. The thesis will further argue that Malle's approach is insufficient due to three reasons: his function argument is very simplifying and therefore troubling; each component of his theory is inconsistent and, finally, closely connected to our common understanding of personhood, which raises new philosophical questions surrounding the basic issue of if and/or when machines can be considered people.

Keywords: robot morality, moral competence, personhood

Ethics in Progress (ISSN 2084-9257). Vol. 8 (2017). No. 2, Art. #6, pp. 69-79.

Doi: 10.14746/eip.2017.2.6