Marija Kušić
Departmant of Psychology
Laboratory for Research of Individual Differences
Faculty of Philosophy, University of Belgrade
Petar Nurkić
Departmant of Philosophy
Institute of Philosophy
Faculty of Philosophy, University of Belgrade

Original Scientific Paper
UDC 004.8: 17.018.21
004.85:[159.9:17

ARTIFICIAL MORALITY: MAKING OF THE ARTIFICIAL MORAL AGENTS

Abstract: *Artificial Morality is a new, emerging interdisciplinary field that centres around the idea of creating artificial moral agents, or AMAs, by implementing moral competence in artificial systems. AMAs are ought to be autonomous agents capable of socially correct judgements and ethically functional behaviour. This request for moral machines comes from the changes in everyday practice, where artificial systems are being frequently used in a variety of situations from home help and elderly care purposes to banking and court algorithms. It is therefore important to create reliable and responsible machines based on the same ethical principles that society demands from people. New challenges in creating such agents appear. There are philosophical questions about a machine's potential to be an agent, or moral agent, in the first place. Then comes the problem of social acceptance of such machines, regardless of their theoretic agency status. As a result of efforts to resolve this problem, there are insinuations of needed additional psychological (emotional and cognitive) competence in cold moral machines. What makes this endeavour of developing AMAs even harder is the complexity of the technical, engineering aspect of their creation. Implementation approaches such as top-down, bottom-up and hybrid approach aim to find the best way of developing fully moral agents, but they encounter their own problems throughout this effort.*

Keywords: *Artificial morality, artificial moral agents, machine learning, moral psychology, hybrid model*

1. Introduction

Artificial Morality is a new interdisciplinary field of research within Moral psychology and Machine engineering (i.e. Robotics). In the last decade, due to technological advances, it has been developing at an exponential rate.[1]

1 The work on this paper has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia through the project *Dynamic Systems in Nature and Society: Philosophical and Empirical Aspects* (No. 179041).

 DOI: 10.5937/BPA1932027K

Synonymously called Machine Ethics, Artificial Morality aims to create self-governing, ethical machines that can "function in an ethically responsible manner", that is, machines capable of making autonomous decisions that are in accordance with the society's norms and moral standards (Anderson & Anderson, 2007, pp 15; Allen, Smith & Wallach, 2005, pp 149). To enable morally functioning machines, Artificial Morality considers different ethical principles or learning procedures that govern human behaviour and enable them to act as moral agents. These governing principles are then algorithmically formalized and implemented in machines, thus creating new artificial moral agents. (Anderson & Anderson, 2007, pp 15; Misselhorn, 2018, pp 161).

Artificial Morality can be classified into the subfields of both computer science (more closely, artificial intelligence) and moral psychology (or moral philosophy), predominately because of its eclectic, interdisciplinary approach (Yampolskiy, 2013, pp 389). As a starting point in creating morally competent agents, it uses the achievements of cognitive science and ethics. The main task, when establishing the basic functioning principles of machines, is the abstraction of elements of human moral reasoning and behaviour (Malle, 2015, pp 243) or formalization of ethical principles into computer algorithms (Yampolskiy, 2013, pp 389).

This paper will try to exhibit the complex structure of the Artificial Morality field by dividing it into three main parts (or problems).[2] The first one is the conceptual problem of machines as moral agents, more closely, the mere possibility of machines being moral agents equivalent to humans. This problem is a philosophical one. It grips the normative nature of the field – modality of moral machines – best conceptualized in the question "can machines be moral agents?". Answering this question requires considering the components of moral agency and realizable ways in which machine behaviour can come close to human behaviour.

The second part considers the descriptive, psychological problem that comes after resolving the previous one, namely, the problem of social acceptance of autonomous machines. Moreover, apart from the machines' ability to "function in an ethically responsible manner", it is important to know whether they are going to be accepted and trusted as such autonomous agents, and what will make them more trustworthy in the eyes of society. The problem of interest is how to make technically functional moral machines to also be socially functional agents. In other words, Artificial Morality also deals with the issue of what characteristics, besides the basic governing principles of moral behaviour, the machines need in order to be more like human agents. The public opinion about the safety of modern technologies,

2 This type of classification cannot be found in the available body of literature, but is a synthesis of our own examination of the field and corresponding extraction of general questions and noteworthy ongoing lines of research.

in this case, moral machines, is an important aspect of making their usage possible. For that reason, the acceptance of machines as integral parts of society is one of the central themes in Artificial Morality. This problem is probably best verbalized as a question of "what is needed for machines to be *perceived* as moral?".

Lastly, the third part will deal with the technical side of moral engineering. It is necessary to decide the way in which these machines will run, that is, what the best approach for implementing moral algorithms is and what kinds of algorithms should be implemented in the first place. Engineers, in cooperation with psychologists and philosophers, are trying to decide which governing ethical principles or machine learning algorithms will give optimal results in real-life conditions and render correct ethical judgements. Moreover, the choice on a conceptual level of a machine's functioning (whether there is going to be a set of basic principles which govern machine behaviour, or if the machine will be able to learn and extract ethical principles from experience and then use them to guide its own moral judgements) implies a specific programming approach, which, then, has its own technical challenges.

Artificial intelligence has been a growing field of work for the past 50 years (Malle, 2015, pp 161), and yet efforts to answer certain questions about morally functioning autonomous AI machines, or artificial moral agents (hereinafter AMAs), had begun only a decade ago (Yampolskiy, 2013, pp 389). A key reason for even stepping into this endeavour of creating AMAs was the rapid development of autonomous machines or decision-making algorithms used in a wide range of everyday situations, from driverless vehicles and elder care robots, to bank intelligent money transfer software (Wallach & Allen, 2009, pp 17; Goodall, 2014, pp 93, Misselhorn, 2018, pp 162).

Consequently, this emerging usage of autonomous systems has increased the number of situations in which they will be put in a decision-making role with a different magnitude of repercussions for the society. Moreover, there are already seemingly paradigmatic examples of the aftermath of judgements in morally oblivious AIs. There have been incidents in which these AIs, as a result of their reasoning process, selected violent videos for children, produced racist tweets or even racially discriminated against convicts on parole when accessing their risk for recidivism (Shank, DeSanti & Maninger, 2019, pp 652). However, there are even more moral decision-making opportunities that we encounter daily. Although they are not as visible as aforementioned scenarios, and thus not used as representative examples for the exigency argument about the implementation of moral decision-making abilities in artificial intelligence, they are vastly frequent and, consequently, more important: For instance, we can briefly focus on the increase of daily usage of automated vehicles and elder care robots. Goodall addresses (2014) the remark that people rarely make moral decisions while driving, and thus

machines shouldn't either, by accentuating the morality of everyday decisions, regardless of how small they may seem (especially when an evaluation about their importance is made based on actualized consequences rather than possibilities). Accordingly, he states that the category of ethical judgements includes cases such as a driver deciding whether to unlawfully speed up so there can be more room for a cyclist on the road (Goodall, 2014, pp 97). Similarly, Wallach and Allen mention the example of medication dispensing robots for the elderly (Wallach & Allen, 2009, pp 15). In its way of completing the task of handing medicine to someone in need, a robot may encounter various obstacles that require ethical judgement about the robot's further behaviour. What if the mentioned obstacle is a child instead of an object? Would the robot's judgement be based on the utility of alternative solutions? Should the robot have a predefined set of preferable actions and rules it follows, or should it be able to learn from experience and examples of correct judgements in order to abstract guiding rules?

There is a shared concern about the possible outcomes of self-guided behaviour in morally oblivious machines (Anderson & Anderson, 2007, pp 16; Goodall, 2014, pp 94; Misselhorn, 2018, pp 162; Yampolskiy, 2013, pp 389; Shank, DeSanti & Maninger, 2019, pp 649; Wallach & Allen, 2009, pp 3), but also a research field that aims to overcome these concerns. This field is called Artificial Morality. Its central approach to preventing possible judgement mistakes of intelligent machines is ensuring that their behaviour towards humans and the environment is ethically acceptable, which is achieved by creating artificial moral agents, AMAs.

2. Modality of AMAs: moral agency of machines

One of the main problems in Artificial Morality is whether machines can be moral agents in the same way that humans are, or at least moral enough to be attributed the characteristic of moral agency. Following the latter thought, there is a discouraging picture of AI's morality in relation to human morality. The public opinion of AIs is more negative than positive, that is, people are distrustful towards intelligent machines and they do not feel at ease about machines making autonomous decisions. In other words, people do not perceive AIs as moral agents, nor do they attribute to them the status of equal members of the society (Bostrom & Yudkowsky, 2014, pp 318). From such state arises a new problem of inequality amongst humans and AIs. Dispositions that can be formularized and implemented in artificial, intelligent machines, which can then simulate them successfully, often get post hoc characterized as not real enough, or even completely disregarded, because of the idea of non-human embodiment (Bostrom & Yudkowsky, 2014, pp 318). Bostrom states (2014) that this kind of rejection of valuable human characteristics, when they are exhibited by machines, emerges from

the recognition of their specialization in a specific domain. For example, AI's ability to play chess or Go better than the champions in these games ceases to be perceived as extraordinary, impressive or valuable because of the awareness that this ability in AI is limited only to this domain (Bostrom & Yudkowsky, 2014, pp 318). This devaluation of human abilities which are not proven as general traits, but instead exist only for a specific purpose, indicates that value is attributed to those characteristics that are applicable in a variety of situations.

In addition to the demand for generalizability of traits, so they can be accepted as human-like, there is also a demand for a less perfect performance (Indurkhya, 2019, pp 108). Perfection and lack of mistakes in a machine's performance of tasks evokes a sense of mannerism and artificiality in humans. Because of the social rejection of AIs manifestation of human dispositions, the efforts for creating widely accepted machines are going in the direction of making their behaviour more human-like. For example, there have been cases of deliberately constructing AIs that make mistakes while performing specific actions such as dancing or drawing (Indurkhya, 2019, pp 109). This issue of public acceptance and required competence for equal and human-like machines will be addressed in the section *Moral competence of machines*. This section will focus on the conditions of moral agency.

There is no universally accepted definition of moral agency in ethics literature. Furthermore, there are frequent disagreements over what constitutes a moral agent (Misselhorn, 2018, pp 163), but despite this division of opinion, there is also a surprising overlap in different understandings of moral status (Misselhorn, 2018, pp 163).

One of these understandings (Misselhorn, 2018, pp 163) highlights two central conditions of moral agency: (1) the subject must be an *agent*, (2) and it must be a *moral* agent.

Agency is then defined through concepts of self-origination and self-reasoning. The concept of self-origination refers to the origin of an agent's action. The agent is here understood as self-originating only if the source of her action is within herself. That means that the action initiators are the internal structure and dispositions of the subject and not external events. The most demanding form of the self-origination concept refers to "the action without any prior cause" except the agent's humour, but a less strict and commonly used form of self-origination is understood as actions that are under the control of the agent, are not solely determined by external stimuli and can be manifested with "greater flexibility that is dependent on the agent" (Misselhorn, 2018, pp 163). In a practical sense, applicable to artificial agents, less demanding criteria of self-origination means that agents are able to interact with the environment, to affect the environment and its own state without the influence of external events, adapt to external conditions or actively change them. The self-reasoning concept considers the capacity to

act for a reason, in other words, the capacity to have a belief in something and a pro-attitude (desire) towards something. The combination of belief and pro-attitude constitutes a reason to act and guide our behaviour.

Furthermore, moral agency is attributed to the agent if her source of action, and reasons for it, come from inner moral reasons. That is, the agent can be a moral agent only if her self-origination and self-reasoning capacities include moral attributes (Misselhorn, 2018, pp 164).

A similar understanding of moral status (Bostrom & Yudkowsky, 2014, pp 321) also extracts two important criteria: sentience and sapience of the agent. Sentience applies to the ability to have qualia, an idiosyncratic phenomenological experience. Qualia is often understood through the capacity to feel pain, but it refers to any kind of emotional or sensory experience. Moreover, it is thought that animals possess, in different degrees, this ability of phenomenological experience. The concept of sapience is understood as a capacity for self-awareness (consciousness) and acting for a reason. This kind of capacity implies higher cognitive structure which can only be found in humans. It can be noticed that sapience incorporates both Misselhorn's conditions of agency (self-origination and self-reasoning) but does not imply moral reasons that she highlights as necessary for moral agency.

It is clear that artificial systems cannot meet the most demanding forms of aforementioned conditions of moral agency, but given that such metaphysical concepts evoke still unresolved debates concerning human agents, there is a justified reason to concentrate on less demanding criteria of moral agency. In the case of the self-originating concept, we should move away from the metaphysical controversy of determinism and initiation without any prior cause. Less demanding criteria understand self-originating agents as agents who can change their environment or their own state without being influenced by external stimuli (Misselhorn, 2018, pp 163). This criterion puts focus on observable elements of situations that guide our conception and attribution of agency. We argue that this form of conclusion about the agency is a justified way of judging about the moral status of machines, given that it appears to be an important aspect of judging about agency when it comes to humans.

When attributing the cause for someone's behaviour, people primarily take into account the situational factors of the event (Kelley & Michela, 1980). Whether the cause of someone's action is going to be attributed to their internal dispositions or to external situational factors, depends on the observable characteristics of a situation. According to the empirically sustained Kelly's Attribution theory, if there is a possible situational explanation for someone's action, the cause of action will be attributed to the external stimuli rather than person's dispositions (Kelley, 1973; Kelley & Michela, 1980). For example, we do not interpret a professor delivering a lecture as her being a talkative person nor do we interpret a waiter's pleasantness as him being a

friendly person. Instead, we exhibit a tendency to explain their behaviour as situationally structured – the professor talks because it is her job to give a lecture, and the waiter is pleasant because his job also depends on his positive attitude. In lack of congruent situational factors, the cause of action will be attributed to inner factors, a person's dispositions (Kelley & Michela, 1980). That means that, for example, we will interpret the waiter's unpleasantness as him being a rude person because there are no relevant, congruent external factors that can overrule attribution to inner, dispositional factors.

The same framework can be applied to the general attribution of agency. If such regularity of attribution of dispositions is noticed when it comes to human actions, then the same logical line should be justifiably followed by a discussion about the machine's actions. Bostrom's principle of ontogeny non-discrimination (Bostrom & Yudkowsky, 2014, pp 323) also states that "if two beings have the same functionality and the same consciousness experience, and differ only in how they came to existence, then they have the same moral status". That means that if artificial systems can act without any situational factors that noticeably influence their actions, they can be attributed with dispositional causes. These inner causes are markers of agency, and if AI systems have the capacity to act according to inner causes and reasons, they will have some status of agency. Additionally, if those reasons are moral reasons, they will have the status of moral agency (Misselhorn, 2018, pp 164).

An interesting view of agency, applicable to AI systems, is provided within the framework of moral psychology. Gray and colleagues (Gray, Young & Waytz, 2012, pp 103) discuss moral agency (i.e. moral judgement) as fundamentally dependent on, and determined by, mind perception. On the basis of extensive research of mind perception, they conclude that people perceive minds through two independent dimensions – the dimension of *experience* and the dimension of *agency*. The experience dimension is analog to Bostrom's concept of sentience, and is understood as the ability for sensation and feelings, while the agency dimension, which can be represented by the concept of sapience, refers to the capacity to act and to intend (Gray, Young & Waytz, 2012, pp 103). These dimensions of mind perception appear to be strongly linked with perception of one's moral status, usually defined through ascriptions of rights and responsibility. Perception of experience is correlated with ascription of rights, that is, with the perceived ability to feel (pain, pleasantness) comes the ability to benefit or suffer. Perception of agency, on the other hand, is correlated with ascription of responsibility, namely, if one is prescribed a higher capacity to act and intend, one could also be attributed more blame or praise (Gray, Young & Waytz, 2012, pp 104).

As Gray and colleagues define it, perception of agency qualifies moral agents and perception of experience qualifies moral patients (Gray, Young & Waytz, 2012, pp 104). As agency and experience (or moral agency and moral patiency) are independent dimensions, there can be entities high in both

dimensions, low in both dimensions, or high in one and low in the other dimension. For instance, adults are perceived as entities that are high in both agency and patiency, and thus can be both responsible (blamed) for their actions and deserve rights (protection) from actions of others. Moreover, AI systems would be perceived as high in agency, which would grant them the status of moral agents, but low in experience, which would deny them the status of moral patients. Essentially, that means that AI systems will always be perceived as entities who act but never receive (feel). Given the omission of perceived capacity for sensation, AIs will not have moral rights, but, given the actualized perception of agency, they will be ascribed to full spectrum of moral responsibility

Moreover, morality is broadly understood as a dyadic interaction between two perceived minds, a moral agent and a moral patient. Gray and colleagues argue that the essence of morality can be captured in this cognitive template of "perceived intentional moral agent and a suffering moral patient", where the presence of moral agent is required but the presence of suffering moral patient can just be imagined (Gray, Young & Waytz, 2012, pp 107).

A dyadic structure of morality recognizes the phenomena of moral typecasting. Moral typecasting refers to the categorization of people either as moral agents or moral patients. Even though this kind of mutually exclusive categorization is apparent within a specific moral context (where a prototypical moral situation revolves around the interaction of a moral agent and a moral patient), moral typecasting suggests a more general categorization – people are usually and consistently seen as *either* moral agents *or* moral patients (Gray, Young & Waytz, 2012, pp 113).

Furthermore, moral typecasting can influence the perception of one's mind, that is, the perception of one's moral status. Those that are categorized as moral agents are ascribed with the capacity for agency and intention, and are given moral responsibility as well, whereas those categorized as moral patients are ascribed with the capacity for experience and are given moral rights (Gray, Young & Waytz, 2012, pp 113). Given that AIs will consistently be found in roles of moral agents, as acting entities with aims and tasks, they will automatically be categorized as moral agents and correspondingly attributed with agency and intention.

Concluding this section, we can see that AIs, by the very fact of fulfilling the roles of agents, can be (and will be) perceived as *moral agents* with a certain level of expected moral responsibility. However, a natural consequence of typecasting AIs as moral agents will create a general, conclusive perception of them only as moral agents, but never moral patients. This puts AIs in an unflattering position. Although they can have moral agency and can be blamed for their actions, they cannot enjoy the status of being moral patients similar to humans or animals, and will thus not be given corresponding moral rights.

3. Moral competence of machines

The discussion about machine morality has so far been focused on their capacity to be moral agents and the problems of defining moral agency. With those tasks ahead come many difficulties about finding one universally accepted definition of moral agency and choosing which of the many understandings of moral agency to follow when deciding about the machine's moral status. Malle, however, proposes a new approach to the problem (Malle, 2015, pp 245): it is more functional to focus on the constituents of human moral competence and use them as orientation guides for creating morally competent machines, instead of focusing on defining moral agency. Understanding the elements of human moral competence can serve as a guide for the making of moral algorithms for machines. This approach ends discussions about machines as moral agents equivalent to humans, and makes room for more fruitful possibilities for designing machines that are competent agents which can perform the needed tasks. They can also have different degrees of competence.

If machines adequately exhibit this moral competence, people can decide on whether they are willing to accept and form social relationships with the machines. Malle's approach of observation of human behaviour as a guideline for designing machines emphasize the relevance of human-like abilities in AIs. Other authors emphasize this approach as a relevant and successful way for accelerating AIs social acceptance as well (Bostrom & Yudkowsky, 2014, pp 317; Malle, 2015, pp 253; Malhotra, Kotwal, Dalal, 2018, pp 4; Indurkhya, 2019, pp 110).

3.1. Human-like competence in machines

Moral competence is an aptitude to successfully perform moral tasks, namely, tasks of moral decision making and moral behaviour (Malle, 2015, pp 255). Furthermore, moral tasks imply the capability of moral cognition that is defined through one's aptitude for judgements of blame and permissibility, recognition of right and wrong, and emotional reactions while performing these moral tasks (Malle, 2015, pp 255). Acceptance of AIs as moral and social agents depends on their ability to meet people's expectations about their moral and social responsibility. The initial idea is that, with performing regular human tasks, AIs will also take on regular human responsibilities. Their capacity to satisfy these expectations, and successfully perform moral tasks, determines in what degree they are perceived as equal members of society (Bostrom & Yudkowsky, 2014, pp 316; Malle, 2015, pp 245).

Central elements of human moral competence, according to Malle, are (1) moral vocabulary, (2) a system of norms, (3) moral cognition and affect, (4) moral decision making and action, (5) and moral communication (Malle, 2015, pp 245). An extensive study of these elements can be found in Malle,

2015, but the highlighting of the importance for machines to demonstrate more human-like characteristics, in order to make them optimal social agents, puts focus on the emotional aspect of human functioning, that is, on the needed emotional aspect of machine functioning.

Moral philosophy and moral psychology dominantly concentrated their research of morality around the study of moral reasoning, thus neglecting moral emotions, up until the 1990s. This leadership of cognitive reasoning in understanding morality was a product of cognitive revolution and the idea that morality, like language, can be expressed through underlying cognitive structures and corresponding transformations (Haidt, 2003, pp 852). Later theories, on the other hand, highlighted the role of emotions, but the most realistic approach to this problem is the comprehension of both moral cognition and moral emotions as backbones of human morality.

The capacity for both moral cognition and moral emotions that humans exhibit lacks in the case of AIs. As discussed in the previous section, machines can be understood as moral agents with an expected moral responsibility, but never as moral patients with related moral rights. AIs are presumably denied moral patiency because they are missing the capacity for qualia. This capacity, besides sensory experiences such as pain, incorporates an emotional life of an entity, that is, a potential for emotional experience. Emotions, or emotional experiences, are reactions to inter– and intra-activity of an organism, with the main function of mobilizing that organism to adaptively deal with such encounters (Ekman, 1999, pp 46). In other words, emotions are mainly responses to threatening and beneficial stimuli with great motivational tendency, attendant facial expressions and phenomenological experience (Haidt, 2003, 854).

The difference between emotions and moral emotions lies in their relation to self (Haidt, 2003, pp 853). According to Haidt, moral emotions are those emotions that are not directed to self but are "linked to the interests or welfare of other people or a society as a whole", whereas other non-moral emotions are always in more direct relation to self and occur as a reaction to influences on the agent. AIs are missing both types of emotional experience. Emotions such as fear, sadness and happiness are mainly categorized as non-moral emotions, given their occurrence in situations directly related to the agent or in situations of less direct relation between the self and the other. Lack of these emotions deprives AIs of moral rights because not only can they not be physically hurt but they are not able to feel emotional pain or gain either, and are thus perceived as entities that do not need to be protected by society, i.e. do not need moral rights. Moreover, the most prototypical moral emotions are elevation, anger, guilt and compassion, as their triggers are usually disinterested stimuli and are easily triggered by tragedies and transgressions of strangers (Haidt, 2003, pp 854). AI's inability to feel guilt if it makes a judgement error and causes tragedies, or to feel compassion or anger

if it encounters tragedy and pain, determines its further behaviour. Given that emotions have strong action tendencies and motivate some kind of response to the eliciting stimuli, AI's emotional oblivion restricts its empathic and helping actions. That influences the social perception of machines' "coldness" and elicits anticipation of their reluctance to help, which again accelerates people's distrust in machines and makes their social acceptance difficult.

Even though machines can be implemented with algorithms of moral acting, and can thus help others and intervene in situations of need, they are still perceived as agents that cannot feel the direct consequences of moral behaviour related to them. Such picture of senseless entities restricts the attribution of moral patiency and makes them humanly distant.

There are, however, other traits that will help AIs to be socially accepted. Bostrom adds several criteria that need to be algorithmically formalized and implemented in machines (Bostrom & Yudkowsky, 2014, pp 317). The central one, to which others may be reduced, is transparency in decision making. The transparency of AIs reasoning process enables its inspection, a matter of significant importance in the possible scenarios of reasoning mistakes or hazards caused by AIs decisions (Bostrom & Yudkowsky, 2014, pp 317). The knowledge of how these intelligent algorithms make their decisions does not only enable the tracking of responsibility (and blame) of machines but also has the purpose of amplifying their social trustworthiness. This openness to investigation removes their "black box" artificial invisibility and excites their similarity to human behaviour. Therefore, it is more than needed to equip AIs with psychologically relevant explanations of their own processes (Indurkhya, 2019, pp 110).

Being equipped with psychologically compelling explanations, such as transparency of processes, may also excite AIs general similarity to human behaviour and their consequential acceptance. Until the wanted level of technical development is reached, and AIs are endowed with senses, further development of machines needs to progress in the direction of psychological openness of their judging processes.

3.2. Responsible AIs

An individual's involvement in society is in social psychology often discussed from the perspective of interactionism. The same perspective can be applied to machines, given the effort put into making them welcome members of society that have the status of moral agents. Interactionism describes identity as a meaning derived from social roles one occupies (Burke & Tully, 1977, pp 883). Social surrounding reacts to the agent with expectations for the agent's behaviour to correspond with her social role, in other words, social surrounding reacts *as if* the agent's identity is appropriate to her role performance. An agent understands that reaction and forms a meaning about her identity that guides her following behaviour (Burke &

Tully, 1977, pp 883). In the case of AI systems, it is important to highlight that these social expectations are derived from the very fact that someone is a social agent (Stouten, DeCremer & Van Dijk, 2006, pp 894).

When AI occupies a certain social role, it will evoke corresponding social expectations about its behaviour and dispositions that are common in humans (Gray, Young & Waytz, 2012, pp 113). AI systems, in this case, need to prove their identity of moral agents by adequately dealing with expected moral tasks. Moral competence, besides moral judgement and emotions, entails conforming to social norms such as the principles of righteousness and equality. If social expectations of honouring these principles are disappointed, people will react with anger, emotional distress and retributive reactions in order to correct the inflicted injustice (Guth, Schmittberger & Schwarze, 1982, pp 384; Stouten, DeCremer & Van Dijk, 2006, pp 895). Because these reactions only appear when someone is perceived as a moral agent, that is, if someone is perceived responsible for their actions and obliged to follow social norms (Gray, Young & Waytz, 2012, pp 113), we can test the moral status of AIs by examining people's reactions to AIs in situations following the violation of social norms.

There has been a new body of experimental literature that grips the above-mentioned problems of AIs' social acceptance. One of the ways to investigate their social status, or at least to scratch the surface of social interaction between humans and machines, is through the Game Theory experiments. These experiments simulate decision-making interactions between players with an aim to reveal and understand the components of their reactions and reasoning (Osburne, 2004, pp 1). Simulated situations of choice often require social choices where subjects can demonstrate their social norms compliance or violation, and reaction to the compliance or violation of others. That is, they can demonstrate their moral competency.

Relevant for these purposes is the bargaining game, called the Ultimatum Game. The simplest and most commonly used version of the Ultimatum Game is the two-player version. This is a bargaining game because one of the players must solve a distribution problem, usually of goods (Guth, Schmittberger & Schwarze, 1982, pp 367). When this player (commonly known as the first player) makes her choice of distribution, she restricts all the possible alternatives of distribution of goods to one proposal (her choice). The other player (the second player) can then only accept or refuse the first player's proposal. In other words, the first player decides on how to distribute the goods (e.g. money) and makes her proposal to the second player who can then only accept or decline. There are no simultaneous moves of players in the Ultimatum Game, but instead, every aspect of the game is successive so that the players can always observe each other's decisions (Guth, Schmittberger & Schwarze, 1982, pp 367; Osburne, 2004, pp 179). That way, every player is, at the same time, always and completely informed of every previous move in the game[3].

3 Such a game is said to have perfect information.

The specificity of the Ultimatum Game is that the bargaining comes in a form of "strategic reactions based on anticipated future events" (Guth, Schmittberger & Schwarze, 1982, pp 368) where the first player takes into account the "fairness" of her proposal to the second player, and the second player takes into account that the alternative option to the first player's proposal, however unbeneficial, is nothing (and is always worse than the proposed distribution). Because of this bargaining aspect, the game is suitable for investigating social norms and moral behaviour. A further variation of the game can be found in the Dictator's Game with perfect information. The Dictator's Game has only one move in which the first player makes a proposal, and the second player has no other option but to accept it.

The question of interest, when it comes to the social status of machines, is whether they will get the same treatment as human players. Given that both the Ultimatum and the Dictator's games are widely used in social interaction researches, there are noticed regularities of choices that people make and emotional reactions to those choices. If these regularities of human behaviour towards each other also manifest in the games with human and machine players, that is, if human players treat machines the same as they treat humans, there can be a more optimistic comprehension of the machines' social status.

One of the robust findings are acts of retribution when one player feels that norms have been deliberately violated by the other player (Guth, Schmittberger & Schwarze, 1982, pp 384; Stouten, DeCremer & Van Dijk, 2006, pp 895). In situations where the first player's proposal exceeds the 70:30 proportion of distributed goods in her favour, the second player would usually decline the offer even though it means that she will end up without anything. This kind of reaction is described as a retributive reaction to what someone understands as injustice (Stouten, DeCremer & Van Dijk, 2006, pp 895). There is also an observed regularity of the prosocial proposals first players commonly make. In most cases their distributions are fairly made, that is, the majority of players distribute goods in an approximately equal share. More specifically, they strive to benefit from their distribution, but to also split the goods according to the fairness norm (Forsythe, Horowitz, Savin & Sefton, 1994, pp 362, Guth, Schmittberger & Schwarze, 1982, pp 384).

Equivalent treatment of machines and humans was demonstrated in one such experiment (Nagataki et al., 2019). No significant difference was found between human and robot status in prosocial and retributive tendencies of participants. All human participants made the same prosocial offer of nearly half of the total amount of money to robots as they did to humans, in both, the Ultimatum and the Dictator's Game. That way, they equally respected the norm of fairness amongst social agents, whether the other agent was human or not. Moreover, the participants rejected "unfair" offers from robots, just as they did from humans, thus demonstrating a will to punish what they

considered unjust behaviour, even at the cost of their own gain. These kinds of equivalent reactions to machines and humans may speak in favour of potentially equal social status between them.

4. Engineering approaches to machine morality

Because the consequences of AIs decisions have an unavoidable impact on humans they need to be treated at least as agents with moral behaviour, regardless of the society's acceptance and the question of their full moral agency (Allen, Smith & Wallach, 2005, pp 149). The very idea behind artificial moral agents (AMAs) is to implement human-like characteristics and learning abilities in them so that they can regulate and monitor their own behaviour, correct themselves and perform better in the future decision-making situations (Wallach & Allen, 2009, pp 15). This is the intersection of work paths of engineers, philosophers, and moral psychologists.

Top-down and bottom-up approaches are two traditional engineering approaches that dictate how different moral principles can be used and algorithmically formalized with the goal of creating AMAs. The third, hybrid approach emerges as a combination of the former two and is insofar the most promising one (Wallach, Allen & Smith, 2007, pp 575: Misselhorn, 2018, pp 166).

4.1. Top-down systems

Top-down approaches are based on fixed normative principles, implemented in AMAs, which are then used as guiding rules of the machine's behaviour. Often called a "rule-based" approach (Allen, Smith & Wallach, 2005, pp 150), top-down models require a general set of moral principles that need to be selected. These principles are then universally obeyed in every situation of moral dilemma and expressed throughout the machine's actions.

One of the first problems with these systems is the selection of moral principles to begin with. There can be an unlimited set of contents from which these main principles can be selected (Allen, Smith & Wallach, 2005, pp 150). Most commonly used moral norms are derived from great ethical theories such as Kantian deontology and utilitarianism, but other frequently named principles are Asimov's laws of robotics, The Ten Commandments or the Torah Commandments (Goodall, 2014, pp 98; Allen, Smith & Wallach, 2005, pp 150; Misselhorn, 2018, pp 166; Yamapolskiy, 2013, pp 389). As it can be seen, these principles can vary dramatically in their generality and number, from three general and unspecified principles in the case of Asimov's laws to the complex computational system needed when it comes to utilitarianism.

Because of their generality, lack of applicability to more domain-specific contexts, and inability to define a concrete set of principles or actions which will guide one's decisions across different contexts and situations, top-down

approaches are severely criticized (Allen, Smith & Wallach, 2005, pp 150). The challenge lies in finding an optimal way of deriving a set of specific rules from the abstract principles. Even though their number is fixed, these rules should be usable in a variety of specific situations.

This approach predominantly uses consequentialist and Kantian theories as starting points in deriving guiding principles for AMAs. Both theories have their own specific problems, but also have a shared one (Allen, Smith & Wallach, 2005, pp 151). The top-down approach based on Kantian deontology encounters the problem of hierarchy of principles, that is, how to submit all its specific principles to one highest principle without contradiction. The other main problem concerns the availability of information, more closely, how AMA should know about the intentions and motives of every agent included in some decision-making situation. Utilitarian AMA faces problems of finding a common value scale for measuring different utilities in various situations and of enormous computational resources needed for even evaluating possible outcomes for every event (Allen, Smith & Wallach, 2005, pp 151). Their shared problem, and a reason for the abandonment of the top-down approach, is the unlikeliness that these algorithms could ever collect and compare every information that they need. This is even more transparent in the cases of consideration of future consequences of actions, instead of focusing on direct and present consequences (Allen, Smith & Wallach, 2005, pp 151).

4.2. Bottom-up systems

The question that imposes itself is how humans restrict their own calculation of continuous external stimuli and predict future consequences since this problem of computational and informational overload is present in the case of their cognitive system as well. Human behaviour is often guided by heuristics and affects decision making (Allen, Smith & Wallach, 2005, pp 151). Moreover, we have the ability to learn from experience and observation. That leads to creating cognitive schemes (scenarios) of plausible events that guide our behaviour when we end up in similar situations (Greene, 2017, pp 69) and, in most cases of decision making, it is what we rely on.

Bottom-up models are based on the abovementioned history of learning, more closely on real data, the experience of correct judgements in decision-making situations from which AMA abstracts moral principles and controls its acts. Bottom-up AMAs do not need an initial set of guiding principles. That means that AMA learns proper moral behaviour while actively participating in their environment (Allen, Smith & Wallach, 2005, pp 151). Bottom-up AMAs can be realized throughout different initial settings and algorithms that determine the type of their learning process. They can simulate learning through trial and error attempts, they can be based on educational learning processes and simulate socialization and the growth of a child, they can simulate evolutionary processes of cognitive and moral growth of an agent,

or they can be based on neural-network processes which associate patterns in the surroundings they encounter (Misselhorn, 2018, pp 166; Wallach, Allen & Smith, 2007, pp 570).

This approach resolves some of the problems that top-down models have by introducing a self-changing and self-improving system. The machines based on Bayesian models can adopt moral rules and change their behaviour when in contact with their social surroundings. They constantly revaluate first guiding principles, as a reaction to new information and experience learning, and verify the consistency of all previously formed rules (Shaw, Stockel, Orr, Lidbetter & Cohen, 2018, pp 73). These machines become self-checking agents capable of human-like adaptation to surroundings.

Bottom-up systems provide more natural and stronger models of moral reasoning that can be an almost ideal approach for creating agents with optimal social functioning and ethically responsible judgements. However, a significant problem of these systems is that they are extremely difficult to develop and usually need a lot of time to evolve into an optimal moral-reasoning autonomous agent (Allen, Smith & Wallach, 2005, pp 151). There is a rising problem of controlling the learning data for AMAs, so cases in which bad data may contribute to their socially unacceptable principles and decisions can be avoided. Other than that, engineers encounter an additional problem of not knowing which principles to use as a guideline in the situations of changed contexts (Wallach, Allen & Smith, 2007, pp 572) and the uncertainty of what will be the evolutionary outcome of a specific AMA (Misselhorn, 2018, pp 167).

4.3. Hybrid systems

Although top-down and bottom-up models represent the most common way of implementation of moral competence in AMAs, their combination is often characterized as necessary for overcoming the specific and general problems that both approaches carry (Misselhorn, 2018, pp 166). Therefore, hybrid systems originate from combining top-down and bottom-up approaches into one "eclectic" model. Hybrid AMAs are implemented with algorithms inspired by both traditional approaches.

Their top-down part is a predefined and fixed set of initial principles that serves as a starting point from which AMAs learn and self-improve. The predefined sets of rules are often not as general as in traditional top-down systems, but are more closely specified to domains in which they are set to be used (Misselhorn, 2018, pp 167). As it was mentioned, hybrid AMAs maintain the ability of self-improvement regardless of their initial moral principles. This ability to learn from experience and adapt is their bottom-up part (Allen, Smith, Wallach, 2005, pp 153), and because of that, their guiding principles often get changed throughout this learning process. Hybrid AMAs, as self-checking agents that are actively involved in the environment, develop

even more specific moral judgements congruent to characteristics of their surroundings (Misselhorn, 2018, 166).

Allen, Wallach and Smith (2005) interpret top-down and bottom-up systems through contrasting the explicit and implicit values and their ways of acquiring. In their description, top-down systems can be understood as explicit values and ethical principles "outside of the entity" that are demanded from a specific cultural milieu, while bottom-up systems are implicit moral values abstracted from practice and experience that then emerge from "within the entity" (Allen, Smith, Wallach, 2005, pp 153). That way, a top-down AMA can be described as AMA of "rights and duties" or "welfare and utility", while bottom-up AMA is an AMA of "practice and experience". As their combination, hybrid AMA is understood as an entity raised in a culture which prescribes its own explicit moral concerns and judgements and requires they be respected, while it (the AMA) still has constant opportunity to discover and learn other values and traits from practice (Wallach, Allen & Smith, 2007, pp 576). That is, AMA is given some kind of parental rules (like those a child is demanded to follow during his or her development) but it also interacts with the environment and through that learns or demonstrates her individual traits.

Because of the above-described hybrid AMA's position between top-down and bottom-up models, Aristotelian virtue ethics is seen as a fruitful framework for hybrid algorithms (Wallach, Allen & Smith, 2007, pp 576). Aristotelian virtuous character resembles ethical principles and initial rules implemented in hybrid AMA because, in both cases, they are initial motivators of one's action and overall behaviour, and are understood as one's features or traits. Furthermore, Aristotelian understanding that moral virtues are learned through experience, and consolidated from constant practice, is equivalent to hybrid AMA's ability to learn from experience and self-improve (Wallach, Allen & Smith, 2007, pp 576).

Hybrid AMA begins to be interpreted as an individual, into which one can instill Aristotelian "good traits", virtues or characteristics, "complex patterns of motivation" (Wallach, Allen & Smith, 2007, pp 577) and personality dispositions that determine our tendencies to act in a certain way. This new line of thinking has the reincarnated problems of top-down models. It is difficult to choose a number and type of virtues machines should have in order to be moral, but importantly, it is extremely difficult to simulate virtues (Wallach, Allen & Smith, 2007, pp 577). Virtues, as dispositions and patterns of motivation, are manifested in one's general way of behaving, i.e. in a variety of different situations. In that sense, one virtue has multiple behavioural expressions and is responsible for various acts. Because of that, AMAs should be able to connect every potential action or judgement to a certain trait, that is, they should be able to "know "every possible manifestation of some virtue so it could adequately practice that virtue in its overall behaviour. Moreover, the traditional problem of constant checking if every chosen action is

congruent with all guiding principles, both specific and higher, requires enormous computational power. Even more computer power is needed for creating a non-contradictory hierarchy of virtues and enabling a changeable AMA to develop and incorporate new virtues in such a non-contradictory way (Wallach, Allen & Smith, 2007, pp 577).

Another approach of implementing Aristotelian virtues in AMAs comes from bottom-up models, specifically, neural networks. The central idea is the development of a virtuous character. Neural network system has access to training data from which it abstracts moral principles, while the further gathering of data is realized in real-life scenarios where network surpasses its previous generalized principles (Wallach, Allen & Smith, 2007, pp 577). However, present perspectives that provide insights into human developmental process still cannot provide adequate frameworks for the learning process of moral virtues when it comes to neural networks (Wallach, Allen & Smith, 2007, pp 578), and for now, this approach remains only a daring idea.

4.3.1. Culturally assimilated AMAs

Allan et al. (2005) discuss the initial set of guiding principles in hybrid AMAs as explicit values of the cultural milieu they are made for. However, this thought is not just a superficial analogy made for better conceptual understanding of the top-down approach. Cultural variation requires the adaptation of machines to specific contexts in which they function. For this to be done, we need to first abstract specific dimensions of morality, and from there conclude which specific dimensions suit which culture.

The Moral Foundation Theory (MFT) provides a conclusive picture of a moral mind "constructed" of a universal set of moral modules, innate foundations which guide the learning process of moral values, norms and rules, and are environmentally sensitive (Graham et al., 2013, pp 10). Haidt and colleges integrate the evolutionary position of innate morality and a constructivist perspective on cultural shaping of values and moral behaviour. They propose that the human moral mind is organized "in advance of experience", that is, it evolved a set of "moral matrices" (shared knowledge) as a tool for solving social problems of a cultural human (Graham et al., 2013, pp 8). These modules are understood as foundational moral instincts that enable the learning of some moral values and behaviours over the other. That way, people are innate with potential for acquiring a set of universal moral norms (foundations). Which of these universal norms will be adopted, which particular values generated and in what degree, will be determined by a specific culture, through one's development process.

Based on the MFT perspective on moral norms we can further discuss which set of initial guiding rules, or moral values, should be implemented in AMAs. MFT proposes five moral foundations that we mark as suitable for AMAs' norms (Graham et al., 2013, pp 12).

(1) The Care/harm foundation represents a functional mechanism that enables association of perceived suffering with actions of nurturing, caring and protection. This foundation is extremely important for machines with highly responsible tasks that revolve around people, such as elderly or children care, but also for machines whose judgement decisions may directly or indirectly influence one's life (automated vehicles). The AMA with values of caring for– and protecting others will presumably be of equal importance across cultures, given that more trustworthy and reliable machines will be valued and demanded– regardless of the individual differences between individualistic and collectivist cultures.

(2) The Fairness/cheating foundation is responsible for being observant to sings of cooperation or cheating amongst others (Graham et al., 2013, pp 13). It generates specific values such as righteousness, fairness, sensitivity to inequality, justice that include retributive behaviour as well as rewarding acts, and so on. Dimensions of Care, Fairness and Sanctity (described in the next paragraph) turn out to be important categories for evaluation of virtuousness in both liberal and conservative groups (Graham et al., 2013, pp 20). Given this invariability to conservativism, the value of fairness would be important for machines to have in different cultures or social groups they are made for. By satisfying social demands for fair judgement, AMAs would prove themselves as responsible and trustworthy members of society.

(3) The Sanctity/degradation foundation relates to sensitivity for puritanism of body and "soul", that is, values and motives for which "people treat their bodies as temples" (Graham et al., 2013, pp 14). It is, as mentioned above, a valuable moral norm for estimation of virtuous character, but it is not invariant to cultural context. The sanctity is extremely important in collectivist and traditional cultures, where AMAs need to adapt to bigger roles of purity and religious concerns in everyday life (Graham et al., 2013, pp 26). That requires implementation of religious beliefs congruent to the market culture of an AMA. Just as it would be preferred that AMAs exhibit dominantly practiced and expressed religious rules, values and norms in Eastern cultures, it would be required that AMAs do not exhibit those same values in a secular society. Moreover, intragroup differences in cultural variation are robustly greater than intergroup differences, that is, these traditional differences are greater within cultures than between themselves (Graham et al., 2013, pp 26). Given regularities such as this, it is better to equip AMAs with values of puritanism and religion according to the tasks they will perform. If the deciding process of the task requires evaluation of such criteria, then its implementation is also needed. These are not just complex tasks, but tasks for which optimal solutions involve cultural knowledge. We already mentioned the example of medication dispensing robots for the elderly (Wallach & Allen, 2009, pp 15). In its way of handing the medicine, a robot may encounter various obstacles that require judgements about whether they are religious objects that need to be carefully avoided or not.

(4) The Loyalty/betrayal foundation highlights the importance of motivational tendencies to exhibit the traits such as agreeableness, fidelity and alliance, because of their significance for forming coalitions and preserving group cohesion (Graham et al., 2013, pp 13). Compliance to this norm makes social functioning, particularly group functioning, possible and thus is an inevitable value for AMAs that are privately owned. Even though loyalty foundation is more connected to the conservative groups (Graham et al., 2013, pp 16), machines need to exhibit alliance tendencies as an acceptance tool.

(5) The Authority/subversion foundation serves as a mechanism for navigating one's behaviour in hierarchical social interactions. It shapes values of obedience and deference (Graham et al., 2013, pp 13) that are, again, more valuable in conservative groups and collectivist cultures than in liberal and more individualistic groups. These values may suit AMAs who have roles of carers and are in direct contact with humans.

All five moral foundations interact with the environment and generate more specific moral values. These foundations are thought to be universal structures of the human moral mind, but their shaping and development is vastly dependent on culture. As we can see, less traditional and liberal groups generate the care/harm and fairness/cheating foundation in greater degree than conservative groups, and more traditional and a conservative environment values the authority, loyalty and sanctity foundations more than liberal groups (Graham et al., 2013, pp 16). That does not exclude some moral norms from certain cultures, but rather priorities values within cultures. Creators of AMAs should thus be sensitive to these cultural moral priorities when making machines for targeted markets, and MFT provides an inclusive and culturally sensitive framework for this kind of deliberation of initial guiding values.

5. Conclusion

This paper had the aim of systemizing the complex, even though new and yet expanding, field of Artificial Morality. Artificial Morality centres around the idea of artificial moral agents (AMAs) which represent self-checking machines able to change and grow while making moral decisions side-by-side with humans. The presented structure of main problems in Artificial Morality originated from the authors themselves. These problems, even though noticeable research obstacles, have never, to our knowledge, been understood as a set of three conceptual problems – philosophical, psychological and a technical one.

In the beginning, we had to inspect the question of moral agency and its theoretical applicability to machines (a philosophical problem). Hopefully, we have given our own insights by proposing a line of thinking about machine

agency similar to the understanding of human agency. In accordance with the less demanding frame of agency, if artificial systems can act without any situational factors that noticeably influence their actions, they can be attributed with dispositional causes just as humans do. Moreover, if AIs exhibit the capacity to act based on their inner causes, dispositions, they will have the agency status. This agency status can be understood as the status of moral agents if those inner causes and reasons were moral reasons.

The problem of social perception and acceptance of AMAs (psychological problem) has a potentially optimistic solution. It was shown that people react to machines in the same way as they do to humans, thus attributing to them social status and responsibility for their actions. Given that the empirical literature on this subject is still limited, these results should be taken with caution, and used more as an implication for further research rather than conclusions. While considering the importance of human-like competence for acceptance of AMAs, a new line of research has emerged. A great deal of effort has been invested in implementing some additional instances such as emotions, consciousness, or other human capacities (like the theory of mind and symbolic understanding), as it is believed that only these competencies can make reliable and fully moral artificial agents. These competencies are indispensable parts of AMAs, not only because of their social acceptance but for their better functionality (Allen, Smith, Wallach, 2005, pp 153). This problem has not been inspected in detail, but its significance for creating functional AMAs will determine following research in the field.

In the end, we have given a brief overlook of the current state of technical advances, possibilities and restrictions in developing a fully functional AMA (technical problem). There are three main approaches to implementation of moral capacities in machines: the top-down, bottom-up and hybrid approach. The first two approaches are traditional and most commonly used systems that are being gradually replaced by their combination, a hybrid model, as they provide only partially functional AMAs. However, hybrid systems prove to be out of the current theoretical and technical reach. Existing frames of learning processes of human moral competence are still incompatible with the mode of neural networks which are integral to hybrid and bottom-up systems. These inspiring ideas in machine learning, even though challenging endeavours, will also determine future efforts in creating AMAs.

We conclude this section with discussion on culturally sensitive hybrid AMAs. From the perspective of the Moral Foundation Theory (Graham et al., 2013), we suggest five moral norms that should be closely evaluated when deciding which initial guiding principles should be implemented in machines. Moreover, we draft some guidelines for acknowledging cultural differences in the valuation of such moral norms but do not offer final solutions. The field of moral psychology, particularly the study of universal moral rules and cultural variation in norms and practices, is still in its developing research stage, and until we have a clearer picture of human morality there will be a limited potential for this kind of extrapolation onto machines.

References

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches". *Ethics and Information Technology*, 7(3), 149–155.

Anderson, Michael & Anderson, Susan. (2007). "Machine Ethics: Creating an Ethical Intelligent Agent". *Ai Magazine*. 28. 15–26.

Bostrom, N. and Yudkowsky, E. (2014) "The Ethics of Artificial Intelligence". In: Frankish, K. and Ramsey, W., Eds., *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, New York, 316–334.

Burke, P. J., & Tully, J. C. (1977). "The Measurement of Role Identity". *Social Forces*, 55(4), 881–897.

Ekman, P. (1999). "Basic emotions". *Handbook of cognition and emotion*, 98(45–60), 16.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). "Fairness in simple bargaining experiments". *Games and Economic behavior*, 6(3), 347–369.

Goodall, N. J. (2014). "Machine Ethics and Automated Vehicles". In *Road vehicle automation.* Springer, Cham, 93–102.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). "Moral foundations theory: The pragmatic validity of moral pluralism". In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.

Gray, K., Young, L., & Waytz, A. (2012). "Mind perception is the essence of morality". *Psychological inquiry*, 23(2), 101–124.

Greene, J. D. (2017). "The rat-a-gorical imperative: Moral intuition and the limits of affective learning". *Cognition*, 167, 66–77.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). "An experimental analysis of ultimatum bargaining". *Journal of Economic Behavior & Organization*, 3(4), 367–388.

Haidt, J. (2003). "The moral emotions". *Handbook of affective sciences*, 11(2003), 852–870.

Indurkhya, B. (2019). "Is morality the last frontier for machines?". *New Ideas in Psychology*, 54, 107–111.

Kelley, H. H. (1973). "The processes of causal attribution". *American Psychologist*, 28(2), 107–128.

Kelley, H. H., & Michela, J. L. (1980). "Attribution Theory and Research". *Annual Review of Psychology*, 31(1), 457–501.

Malhotra, C., Kotwal, V., & Dalal, S. (2018, November). "Ethical Framework for Machine Learning". In 2018 ITU *Kaleidoscope: Machine Learning for a 5G Future (ITU K)* (pp. 1–8). IEEE.

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.

Misselhorn, C. (2018). "Artificial Morality. Concepts, Issues and Challenges". *Society*, 55(2), 161–169.

Nagataki, S., Ohira, H., Kashiwabata, T., Konno, T., Hashimoto, T., Miura, T., ... & Kubota, S. I. (2019, June). "Can Morality Be Ascribed to Robot?". In *Proceedings of the XX International Conference on Human Computer Interaction* (p. 44). ACM.

Osborne, M. J. (2004). *An introduction to game theory* (Vol. 3, No. 3). New York: Oxford university press.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). "When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions". *Information, Communication & Society*, 22(5), 648–663.

Shaw, N. P., Stöckel, A., Orr, R. W., Lidbetter, T. F., & Cohen, R. (2018, March). "Towards provably moral AI agents in bottom-up learning frameworks". In *2018 AAAI Spring Symposium Series*.

Stouten, J., De Cremer, D., & van Dijk, E. (2006). "Violating Equality in Social Dilemmas: Emotional and Retributive Reactions as a Function of Trust, Attribution, and Honesty". *Personality and Social Psychology Bulletin*, 32(7), 894–906.

Wallach, W., Allen, C., & Smit, I. (2008). "Machine morality: bottom-up and top-down approaches for modelling human moral faculties". *Ai & Society*, 22(4), 565–582.

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

Yampolskiy, R. V. (2013). "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach". *Philosophy and Theory of Artificial Intelligence*, 389–396.