Bachelor Degree Project

UNIVERSITY OF SKÖVDE

MORAL MACHINES

The Neural Correlates of Moral Judgment and its Importance for the Implementation of Artificial Moral Agency

Bachelor Degree Project in Cognitive Neuroscience
Basic level 22.5 ECTS
Spring term 2020

Kristin Winnerheim

Supervisor: Stefan Berglund
Examiner: Andreas Kalckert

Abstract

Society and technology are advancing, in which morality is being artificially implemented into machines, often known as artificial moral agency. Along with this implementation, knowledge about the underpinnings of morality, such as the neurocognitive and ethical basis are an important matter. Human moral judgment has been speculated to be a function for survival, as it favors altruism and prosocial behavior. The neural correlates of moral judgment stem from several structures of the human brain that control cognitive and affective functions such as decision making, cognitive control, theory of mind (ToM) and empathy. In relation to these, the ventromedial prefrontal cortex (vmPFC) has been widely linked to moral behavior such as ToM and moral judgment. The anterior cingulate cortex (ACC) has been linked to regulation of conflict and the dorsolateral prefrontal cortex (dlPFC) to remain cognitive control, which both have strong correlations to moral behavior. Damage to areas such as the vmPFC and ACC have demonstrated abnormal response to guilt, ToM, empathy, risky behavior as well as sociopathic tendencies, which emphasizes the importance of these structures for human morality. By investigating research in the fields of cognitive neuroscience, moral psychology and machine ethics, this thesis aims at seeking out the importance of cognitive neuroscience for the development of artificial moral agency and to furthermore discuss the necessity of emotions in artificial moral agents, which naturally lack the affective part of moral judgment. Lastly, this thesis will cover some of the main dilemmas with this integration and some future implications.

Keywords: neural correlates, moral judgment, deficits, machine ethics, artificial moral agents, moral dilemmas

Table of Contents

1. Introduction

Some people say it is a gut feeling, an innate function that guides our behavior in different settings. Some say it is connected to religion, whereas some suggest it is a universal system independent of theological beliefs. Despite what you want to call it, most humans have a moral sense, principles to act according to or a cognitive reason system that guides our decision making. Every day we must face both trivial and complex decisions, from standing in the grocery store and choosing the best shampoo or decide what to eat for dinner. According to research in decision making, more options do not make it easier for us to choose, and it often results in not choosing the initial item you wanted in the first place or not choosing anything at all (Iyengar & Lepper, 2000). In terms of complex decisions that we need to think through thoroughly, these involve higher cognitive reasoning, not because one has many options, rather because our moral compass does not point in one single direction and moral dilemmas occur. Two of the most well-known moral dilemmas used in moral cognitive research are the trolley- and footbridge dilemma. Both involving a rushing trolley heading toward people who have been tied up on a rail. In the trolley dilemma, one can choose to either save five people from the incoming trolley or one, by pulling a lever to steer the train onto a different path in which only one person lies. Most people who have participated in this task, usually choose to sacrifice one person over five. In the footbridge dilemma, the participant has the option to save those five people lying tied-up on the rail by pushing a person off a bridge to stop the trolley. In this case, people more often choose not to push the person off the bridge and sacrifice the other people. But what are the differences? Research on this topic usually concludes in that pushing a person to stop a trolley from killing people in opposite to pulling a lever to change the direction of the trolley, feels more personal. Personal dilemmas evoke an emotional response that navigates our moral compass into not sacrificing the person on the bridge (Greene, 2014[c]). These findings suggest that emotions

may be highly involved in our moral judgment. However, the trolley- and footbridge dilemmas are just hypothetical, but real-life dilemmas are starting to occur in line with the development of autonomous agents and vehicles.

As society and the development of artificial moral agents increase in industry and in everyday life, moral cognition and machine ethics becomes increasingly relevant (Cervantes, Rodríguez, López, Ramos, & Robles, 2016; Wallach, 2008). The fundamental basis for understanding how to develop advanced machines that can make decisions for humans, is consistent with understanding of the human brain and its structures and functions laying ground for morality, as neuroscience provide knowledge for the development of artificial intelligence (Cervantes et al., 2020). Moral cognitive components such as moral judgment, cognitive control, theory of mind and empathy, are a big part of our moral system (Bzdok et al., 2012). By investigating the functions of human morality, this knowledge will perhaps contribute to the development of machine morality. In contrast, moral judgment involves a complex kind of decision making, as one needs to take values, norms, and principles into account. With the exception of complex rules, implementation of machine morality might be one of the hardest tasks to solve in regards to having both rational and affective decision making in mind (Wallach, Franklin, & Allen, 2010), something that is going to be further discussed. The field of moral cognition is developing rapidly (Greene, 2015) and its underpinnings in moral psychology and cognitive neuroscience, can increase the understanding of these functions for machine intelligence as they compose of several vital functions for human morality.

In this thesis, I will present moral cognition and moral judgment and its neural correlates from a cognitive neuroscientific perspective, to further discuss its importance for machine ethics and artificial moral agents. To additionally, discuss what engineering and machine ethics can acquire from research in neuroscience and human morality when

implementing morality into machines. Along with the development of more advanced machines and computers, there is a need to understand the basis of moral cognitive components, as the greater the freedom of an artificial agent; the bigger the moral responsibility (Picard, 1997).

1.1 Structure and Method

The following essay is structured around the aim of this thesis, initially presenting the theoretical background of the field of moral cognition, including a historical and evolutionary perspective on the field. To further on, present the neural correlates of moral cognition and deficits on modalities responsible for moral cognitive components such as moral judgment, theory of mind and empathy. Lastly, this thesis will include a neuroscientific perspective on the implementation of moral cognitive components in machines and artificial moral agents.

Current and previous knowledge within the field of cognitive neuroscience, moral psychology and machine ethics will be included. To provide an overview of the current knowledge within these fields and reach the aim of this thesis, an extensive literature search will be carried out, using search engines such as LibSearch and Google Scholar, and databases such as Scopus. The selected research will be mainly between 2010-2020. Quality and credibility will be considered regarding publications, journals, and authors, to be able to provide the most relevant research for this thesis.

2. Moral Cognition and its Underpinnings

2.1 Ethics

Moral cognition has grown to be a hot area of research the past years, along with our understanding of the components of moral cognition and its functions. As Greene (2015) argues, "…the field of moral cognition does not study a distinctive set of cognitive processes. Instead, it studies a set of psychological phenomena bound together by a common function"

(p. 40). Two of the perhaps most common ethical theories that are discussed in relation to moral cognition and moral dilemmas, are utilitarianism (consequentialism) and Kantian ethics (deontology). Two philosophical schools within normative ethics that explain how to act morally, but completely distinct in their ideas. Deontology is built upon the meaning and value of an action, and the duty of acting in a certain way, as often known as "duty-ethics". Immanuel Kant argued that the value of acting in according to the "will", weighed more than the outcome or consequence of an action. According to the categorical imperative, as Kant suggested, one should only act according to a maxim, if that maxim at the same time can become a universal law (Kant, 1785/2011). In contrast to deontology and its basis in the will, utilitarian ethics is built upon the consequences of one's actions. Weighing cost-benefit of an action determines whether it is immoral or not. And by having the cost-benefit in mind, utilitarians such as one John Stuart Mill, proposed that a moral action should promote happiness and minimize suffering to the greatest extent (Mill & Corporation, 2011). An example of the scenario where one must weigh cost and benefit are the famous trolley dilemma, where most people choose to save those five people lying on the rail by sacrificing another person to steer away the trolley (Petrinovich, O'Neill, & Jorgensen, 1993). Haidt (2008) explains the moral system as the regulation of selfishness and where the moral system intertwines with functions such as values and psychological mechanisms built to favor social life, an approach that includes a genetic and evolutionary perspective on moral, but also a cultural aspect (Haidt, 2008).

One of many discussions regarding the deontological and utilitarian moral system is which one to interpret as most ethical, and which one that considers both emotional and rational aspects in human morality. There is wide disagreement whether a pure utilitarian approach can be considered as more ethical than a deontological approach and vice versa. This will be discussed further in the discussion section, but to introduce the complexity of

human morality, the problematic aspects need to be considered. For instance, a utilitarian approach has often been associated to a more rational moral system. Some also agree that a utilitarian approach can appear as unemphatic, as its been associated to lowered empathic involvement in moral dilemmas, and in some cases in people with psychopathic traits (Takamatsu, 2018; Takamatsu & Takai, 2019). In opposition to this approach, others argue that sacrificial utilitarian judgments do not correlate with a decreased affective concern. As proposed, utilitarian inclination is not necessarily *increased*, more rather that the deontological inclination is *decreased*, which result in participants that more often make decisions from a utilitarian perspective (Li et al., 2020). Conway and colleagues (2018) argue from the same point of view, that antisocial traits display reduced deontological behavior, which indirectly favors the utilitarian approach due to the lack of deontological influence during moral decisions. This make it easier for participants with antisocial tendencies to sacrifice people. The authors also argue that participants who choose the utilitarian/cost-benefit option in trolley-dilemma is due to the concern of other people and that they want to minimize harm, which indeed displays moral behavior (Conway, Goldstein-Greenwood, Polacek, & Greene, 2018). These disagreements are important in matter of the artificial implementation of morality, as it propose that there may not be a single correct or incorrect moral system, rather that we may need a combination of the two, or a different approach to the problem.

2.2 Morality and Evolution

Morality and its function in evolution, is something that has even confounded Darwin. How does natural selection favor morality? What is morality's role in evolution? According to Greene (2014[a], p. 21), morality can be explained as "…a set of psychological adaptations that allow otherwise selfish individuals to reap the benefits of cooperation". Cooperation and altruism seem to be one of the key components when defining morality.

Without cooperation, there is no competition. And competition has its underpinnings in natural selection. By working toward a common goal, a unified group has a better chance to outrival weaker groups. This altruistic behavior involving cooperation, derives from competition (Greene, 2014[a]). The idea of altruism, however, is limited within the group, and does not function in the same way toward out-groups if the groups do not have a common goal, laying ground for an inter-group bias (Fujino et al., 2020). By working together as a unit, it brings a higher chance of survival, from an evolutionary perspective. This tendency to co-operate, is also known as social regularity, the most common moral behavior in non-human species, and the most basic moral behavior. Reciprocal and prosocial behavior has been correlated with activation in the saliency network, including the insula and anterior cingulate cortex. Moreover, reciprocity has displayed overlapping activation with the default mode network (DMN), involving the medial prefrontal cortex (mPFC), medial temporal lobe (MTL), temporo-parietal junction (TPJ), posterior cingulate cortex (PCC) and the precuneus. Additional limbic activation has also been found, particularly in the amygdala and ventral striatum, which suggests both strong cognitive- and emotional involvement in reciprocity (Cáceda et al., 2017). Reciprocal behavior has also been observed in primates, but their awareness of reciprocal behavior is questionable. One can speculate that these evolutionary pressures led to reciprocal behavior in humans, and eventually morality (Decety & Wheatley, 2015). The evolutionary perspective on morality can also be explained as a bottom-up theory of ethics. Namely, because evolution triggers adaptation and selection, and evolution shape an environment that guide individuals to make moral decisions. By acting morally, individuals can further reap the benefits of their moral actions. Such as in teamwork and altruistic behavior, or learning from experience, as also suggested by Aristotle (Wallach, 2008).

2.3 Theories in Decision Making

Similar to the functions of a digital camera, Greene (2014[c]) argues that the brain has similar functions when it comes to decision making, meaning that the manual and automatic mode in a modern camera is similar to how the brain works. The instant, fast and automatic "point and shoot" mode, has been speculated to be much similar to the emotional response in our brain when we make decisions based upon our emotions, also known as system 1. Whereas the manual mode, system 2, where one can adjust the settings such as exposure and shutter speed, can be compared to the "manual", slower and more deliberate system during decision making (Greene, 2014[c]), such as when weighing decisions. Greene describes the automatic mode as typically non-conscious, that some emotions can be experienced during such events, but most of them appear without us even knowing. The automatic and emotional response has been associated with activation in the ventromedial prefrontal cortex (vmPFC), whereas the manual response that often appear *after* the automatic, involve activation in the dorsolateral prefrontal cortex (dlPFC), as suggested to be more rational and driven by conscious actions. The dlPFC has been suggested to aid our complex decisions, as it guides our behavior towards a specific goal, while the vmPFC which is interconnected to the amygdala, is associated to direct emotional responding, and a "gut feeling". If judging the theory from a philosophical perspective, the automatic setting incline towards an intuitional explanation, where the manual response in our brain can be interpreted as more utilitarian (Greene, 2014[c]). However, is the brain purely fast or slow when it comes to decision making? Greene argues in the light of the dual-process theory, that it is important to acknowledge the misleading aspects of the camera metaphor. For instance, the camera is either turned on or off, in automatic mode or in manual mode. The human brain is always on, especially our automatic, emotional responses which cannot be turned off or neglected (Greene, 2014[a]). Białek and De Neys (2017) proposed a hybrid dual process model, which

they argue that there may not just be a purely fast and intuitional system 1 and a slower, deliberate system 2. Rather, the first system which regard emotions in decision making, are combined with the utilitarian, more deliberate second system, and these systems or responses co-activate during moral dilemmas, instead of being separate from each other. The authors propose a combined system 1 response to moral dilemmas, both intuitional and deliberate. The system 2 is simply unnecessary because one can take on both a consequentialist and intuitional/emotional view on a moral dilemma from a system 1 perspective (Białek & De Neys, 2017). However, the theory itself is quite young, and further research on this alternative dual-process theory is needed. The utilitarian reasoning system in the dual-process theory is often described as slower. However, Trémolière and Bonnefon (2014) found that when increasing kill-save ratios from saving 5 over 1 in moral dilemmas, to saving 500 over 1, the utilitarian response got more efficient and effortless. This suggests that nevertheless the utilitarian response is slower, it can be intuitional and fast as well, contradicting the dual-process theory (Trémolière & Bonnefon, 2014).

3. The Neural Correlates of Moral Cognition

The neural correlates of moral cognition are often studied by using functional magnetic resonance imaging (fMRI). The fMRI uses a blood oxygen level-dependent (BOLD) signal to detect changes of blood flow in the brain, displaying how e.g. different types of tasks or moral dilemmas would change the pattern of blood flow in the brain (Garrigan, Adlam, & Langdon, 2016). Moral cognition, includes both rational and affective functions of the human brain responsible for moral judgment, theory of mind and empathy (Bzdok et al., 2012). These functions have relevance for the field of artificial intelligence and engineering, as the neural and cognitive basis of moral judgment and human moral cognition can facilitate the development of artificial moral cognition.

There are different complex processes that are involved in our moral judgment. A function that our moral judgment highly relies on to function properly, is cognitive control (Jackson, Kleitman, Stankov, & Howie, 2017). This executive function enables us to suppress irrelevant stimuli and attain goal-directed behavior, but also to guide us when we must change our plans due to unexpected events (Norman & Shallice, 1980). Cognitive control and conflict regulation has been associated with activation in areas such as the dorsolateral prefrontal cortex (dlPFC), and the anterior cingulate cortex (ACC) during complex moral dilemmas (Greene, 2014[b]; Greene, Nystrom, Engell, Darley, & Cohen, 2004). These areas are also associated to the manual, reason-system in the dual-process theory (Greene, 2014[c]). Greene and colleagues (2004) found in their study of decision making and cognitive control, that the dlPFC influence cognitive decision making, more explicitly as utilitarian judgment where higher cognitive functions need to be active. According to the authors, the dlPFC's specific role in decision making is to facilitate cognitive decisions that require a higher capacity, by sustaining control and regulating emotions when facing difficult decisions. Complex decision making was also correlated to increased activity in the ACC during activity in the dlPFC, and the authors hypothesized that the ACC initiated the activity in the dlPFC to regulate and control decision making processes. In a meta-analysis on moral decision making, Garrigan and colleagues (2016) investigated whether moral response decision tasks (MRDs) versus moral evaluation tasks (MEs) displayed differences in activation. The moral response decision tasks primarily involved whether how a participant would make a subjective decision in a moral dilemma, whereas moral evaluation tasks involved deciding if an action of another person is morally appropriate or not. Results overlapped between the MRDs and MEs showing activation of the left middle temporal gyrus, cingulate gyrus and medial frontal gyrus when making decisions for oneself but also when evaluating other people's decisions. Additional activation of the right precuneus and the middle temporal gyrus during MRDs can

be interpreted as a difference in making a moral judgment for oneself or judging someone else's moral decisions, as the former, possibly involving more abstract reasoning (Garrigan et al., 2016).

Current research argue to what extent other components of moral cognition such as moral judgment, theory of mind (ToM) and empathy, are driven by the same structures and functions, as several modalities seem to be interconnected. Bzdok and colleagues (2012) found overlapping activation of theory of mind and moral decision making in the (bilateral) ventromedial prefrontal cortex as well as the dorsomedial prefrontal cortex (dmPFC). Along with prefrontal activation, the temporo-parietal junction (TPJ) as well as the right middle temporal gyrus was activated during ToM and moral decision making (Bzdok et al., 2012). Eres, Winnifred and Molenberghs (2018) explored which modalities regulate different aspects of morality based on previous fMRI studies. They found consistent activation in the areas highly related to moral judgment such as dmPFC, vmPFC, the left amygdala and precuneus. These areas showed co-activation during several moral tasks, but the vmPFC showed activation both during affective and cognitive decision making, more specifically when subjects empathized with victims or people in distress.

In summary, the dlPFC and ACC seem to be involved in deliberate moral decision making in utilitarian judgment (Greene, Nystrom, Engell, Darley, & Cohen, 2004), where conflicting stimuli also require cognitive control and goal-directed behavior. However, moral cognition involves more than decision making. Theory of mind is a cognitive function that enables us to "mentalize" or infer other peoples' intentions, actions and emotions. Areas related to ToM have consistently showed activation in the PCC, mPFC and the TPJ (Alcalá-López, Vogeley, Binkofski, & Bzdok, 2019). In contrast to empathy, which refers to when you adopt someone else's emotions, ToM involve the cognitive aspect of understanding someone else's behavior, intentions and emotions (Bzdok et al., 2012).

3.1 The Role of Emotion in Moral Judgment

Emotions and its influence during moral judgment have been positively correlated with vmPFC activity, compared to more rational decision making that is associated with the dlPFC (Bzdok et al., 2012; Greene, 2014[c]). Moreover, paralimbic activation in the amygdala and insula have been associated with emotion elicitation in decision making, and additionally the ACC, PCC, inferior frontal gyrus, middle temporal gyrus, superior temporal gyrus and the TPJ. These areas seem to be involved in both simple and more complex decision making (White et al., 2017). Research on amygdala and its role in affective decision-making is more integrated into the research on moral judgment. Findings indicate that the amygdala plays an important role in decision making, with the area activated during utilitarian judgment and specifically when evaluating utilitarian options as being more immorally correct and when the options appear as emotionally repelling (Shenhav & Greene, 2014). Shenhav and Greene (2014) also found that co-activation between the amygdala and the vmPFC was stronger when making decisions purely based on one's emotions, and the lowest in activation when keeping decisions as strictly utilitarian as possible (Shenhav & Greene, 2014), which suggest a strong interconnection between the amygdala and the vmPFC. Research also shows that affective conditioning before decision making in moral dilemmas affect the outcome of the choice. For instance, presenting participants with a neutral or positive stimulus before decision making, increased the probability to push the person off the bridge in the footbridge dilemma, suggesting a change in sacrificial judgment depending on one's mood. The active or passive decision, whether there was appropriate to push the person (active) or more appropriate not to push (passive) the person off the bridge, affected participants to push the person during positive moods and when it was appropriate (active). Additionally, when participants were in a negative mood and passive frame (not appropriate to push), they were more likely to sacrifice the person on the bridge (Pastötter, Gleixner, Neuhauser, & Bäuml,

2013). This suggests that emotions have a strong influence on human decision making, to further, display the inconsistency of human moral judgment. Similarly as in Pastötter and colleagues' study, Shukla and colleagues (2019) found that mood induction prior to performing the Iowa Gambling Task, changed participants bias toward the positive and negative card decks. Namely, participants that had been induced with positive stimulus through images and sounds, were inclined toward choosing the "correct" or positive card deck, resulting in increased wins. In contrast, participants that had been affected by negative stimuli prior to the task, showed less bias toward the positive decks. These findings contribute to the notion that decision-making is highly influenced by emotions (Shukla et al., 2019).

3.2 Deficits in Moral Judgment

The absence of emotions and morality during decision making is something that has been linked to psychopathic traits. Research in criminal psychopaths using fMRI, has shown that the amygdala, ACC, PCC, ventral striatum and inferior frontal gyrus have been associated with decreased activation during affective stimuli (Kiehl et al., 2001), which indicate the role of the amygdala for emotional processing together with the prefrontal cortex. Additionally, research on incarcerated individuals with psychopathy have also revealed, besides than reduced activity in ACC, a decreased activation in the TPJ and dlPFC during controversial moral dilemmas, suggesting that psychopaths have an altered moral judgment compared to healthy controls (Fede et al., 2016). Dishonest decision making has also been linked to reduced ACC activation in a study on incarcerated psychopaths. Higher scores in psychopathic traits also predicted the reduced ACC activation, and shorter response times to dishonest decisions seemed to be mediated by the lowered ACC activity. This suggest that the ACC might be crucial for regulating conflict during decision tasks, as the decreased ACC activity made participants to choose dishonestly with a short time delay (Abe, Greene, & Kiehl, 2018). Further on, it appears that the vmPFC play a vital role in moral judgment, both

in intentional (planned and accurate) and unintentional (implicit and spontaneous) moral judgment. Cameron and colleagues (2018) found that patients with damage to the vmPFC had difficulties with unintentional judgment, impaired intentional judgment as well as reduced ability to follow instructions regarding intentional judgment tasks compared to the control group. Intentional judgment has been found to be an important factor in regulating affect during decision-making, meaning that a lesion to this area could cause problems in controlling impulses during decision-making (Cameron et al., 2018).

Damage to the vmPFC, not only affect intentional and unintentional moral judgment, it has been demonstrated that damages to the vmPFC affected individuals' decision making. Participants made more risky decisions, and these occurred more often when they performed The Iowa Gambling Task (Bechara, Tranel, Damasio, & Damasio, 1996). In the famous case of Phineas Gage, damage to the very same area resulted in a sociopathic-like behavior where emotions seemed absent. As one of the most investigated cases in the history of medicine and neuroscience, Gage was in 1848 bizarrely injured at a construction site in New England for the Rutland and Burlington Railroad, where he worked as a construction foreman. After a distraction, Gage came too close to the blast area where they detonated rocks to drill holes. As a result, a 109-cm-long and 3-cm-thick tamping iron penetrated his skull, which caused a lesion to his prefrontal cortex. Remarkably, he was still conscious and able to stand up after the accident, but his personality came to be changed forever as his ability to process emotions and make reasonable decisions was vastly impaired. (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994). Similarly, patients with frontotemporal dementia (FTD) have displayed similar behavior as Phineas Gage. In a series of case studies, patients with FTD exhibited sociopathic behavior accompanied by pedophilic tendencies, theft, and sexual harassments. Their empathy seemed absent, and even though the patients knew what they have done was immoral and illegal, they exhibited no signs of guilt or restraint to prevent

their immoral behavior from appearing, suggesting a reduced empathic concern (Mendez, 2010). Another case study with an FTD-patient also displayed reduced moral behavior, along with difficulties to understand the consequences of acting immoral. The patient left his young grandchildren all by themselves at a gathering, suggesting that they would find their way back home by themselves. In this case, brain imaging revealed that the patient had atrophy in the right and medial orbitofrontal cortex, along with anterior temporal areas. This resulted in decreased consequential thinking, and a lack of awareness of why his actions were wrong and immoral (Narvid et al., 2009). These examples display some important distinctions in moral cognition. A lack of empathy may result in a person who cannot feel compassion and guilt, but they do understand why their actions are wrong. In contrast, people with ToM difficulties cannot seem to grasp what they did was wrong, nor feel remorse. Deficits in theory of mind also are speculated to be one of the main contributors to decreased understanding of other peoples' emotions in autism (Narvid et al., 2009).

Regarding vmPFC damage, it appears that emotions play an important role in guiding our decision making, and without them, it is more likely to make poor decisions (Greene, 2014[b]). This is a question highly relevant for artificial intelligence and how this applies to machines, that lack the affective component of moral judgment.

4. Machine Ethics and Artificial Moral Agents

To understand the terminology in the fields outside of cognitive neuroscience, I will provide a brief description of the key concepts used in this thesis. The field of Machine Ethics is a subfield of computer ethics, but in contrast to computer ethics which involve moral issues in regards to the usage of computers, machine ethics is a field which focuses on computers or artificial moral agents and their ethical behavior toward humans (Anderson & Anderson, 2007). The goal of machine ethics is to implement an ethical framework into

machines, such as a deontological or utilitarian moral system (Tonkens, 2009). The term "Artificial Moral Agents" (AMA) was originally coined by Allen et al. (2000), and refers to artificial moral agents in which one can implement values, rules and duties, along with ethical guidelines and principles. Isaac Asimov (1950) was one of the pioneers in developing moral rules for robots, with the aim at preventing robots from acting unethical toward humans (Cervantes et al., 2016; Wallach, 2008). He formulated three rules that would serve as a fundamental basis for the development of robots; that they never were allowed to harm humans or allowing humans getting injured; that they would obey humans as long as that rule did not compete with the first one, and lastly; protect itself as long as it does not conflict with the first and second rule. Nevertheless these rules seem clear in theory, obeying these laws would require an artificial moral agent that is able to decide when it is appropriate to make exceptions of these rules, as well as having an understanding of human intentions (Wallach, 2008).

5. The Integration of Artificial Moral Agency and Cognitive Neuroscience

Since its birth, artificial intelligence has competed with human intelligence, with the goal to reach an equally intelligent system such as the human brain. However, it appears that we are still on a pursuit towards flawless artificial intelligence, machines that can make complex decisions and quickly adapt to different scenarios such as humans. The development of complex intelligence has reached a bottleneck when it comes to moral cognition such as decision making and cognitive control. As Wallach (2008) proposes, engineers and philosophers must combine their expertise when designing artificial moral agents, having both the philosophical values and dilemmas in mind as well as knowing the limits of technology. But due to the lack of adaptability of pre-existing computer models like the Turing machine model (Turing, 1950) and Von Neumann's model of computational architecture, the ability for problem-solving gets limited (Wu, 2019). One common procedure to test a computer's

ability to conversate with a human, is performing the Turing test. The test is built upon a conversation between a human and a computer, examining whether the human can evaluate if the conversation is with another human that for example, stands behind a curtain, or if the conversation is with a computer. The complexity of the computer replies, determines whether the human participant will reveal the computer (Picard, 1997). The main difference between the Turing machine model and Von Neumann's computer architecture is that the Turing machine processes and manipulates input from a series of symbols as when conversating with a human for example. In contrast, the Von Neumann's architecture is limited to stored memory, which performs around its pre-programmed commands (Wu, 2019). The main issue with these models, however, is that the Turing machine is restricted to a limited number of rules for the processing of symbols, which decreases the function of problem solving. Additionally, the Von Neumann's architecture is restricted to a pre-set program which cannot change in case of external or on-demand changes (Wu, 2019). Regarding the human brain, we can change our behavior instantly when a situation changes and are not restricted by rules or pre-programmed ideas. However, human decision making is not that consistent, and we easily change opinions especially when driven by our emotions. Technology is learning more about implementing evolutionary algorithms into machines. In other words, making systems able to select and prioritize tasks to reach a goal, just as humans have done to evolve. Sometimes referred to as bottom-up strategies. Bottom up architectures in AMAs build its representation of the world, and its input from the current environment. Top-down strategies can be explained as when trying to solve the nature of the problem, deciding what is moral and what is not, for example implementing moral rules such as Kant's categorical imperative or the Ten Commandments in an artificial moral agent (Wallach, 2008). Moor (2006) proposed another view on how to categorize AMAs other than through bottom-up or top-down strategies. He divided AMAs into different classifications depending on their level of autonomy and

understanding of ethical behavior. For example, an Implicit ethical agent cannot interpret what is good or what is bad behavior, but it can act morally and avoid immoral behavior because it has preprogrammed ethical rules. Explicit ethical agents, however, are enabled to act on several different ethical systems such as deontological or utilitarian rules, among others. Lastly, full ethical agents are agents that are closest to humans, autonomous, with a capability of intentions, desires, and a free will. These agents can make complex moral decisions and know how their actions can be justified. Having a classification of moral agents increase the understanding of the moral autonomy of the agent, however, there are still only humans that can be seen as full ethical agents in the current stage of research. The taxonomy does not include how to implement it technically, only the theory behind it (Cervantes et al., 2020).

Along with the development of methods in neuroscience, artificial intelligence can more easily reap the benefits of research in neuroscience for the development of artificial moral agents. The human brain appears to be the most natural choice of inspiration for the development of AI-technology (Cervantes et al., 2020; Wu, 2019), as it is a system that is capable of complex moral decision making in terms of both reason and emotion (Wallach et al., 2010). It is a challenge to implement conscious emotions in moral judgment in machines, but it may be possible for machines to interpret human emotions according to affective computing. By programming machines that can interpret human social cues such as facial expressions, vocal changes and body language, robots and humans can more easily interact with each other. For engineers, the main issue does not lie within programming the robot, however. Rather, to understand how humans perceive the machines we must be able to interact with, and in reverse, be able to program robots to interpret our social cues (Wallach, 2008). This strengthens the importance of understanding the cognitive and neuroscientific aspects, as emotions are the most problematic component to implement in machines.

Nevertheless, this might be the most important component as emotional input during human moral judgment can facilitate moral behavior (Mendez, 2010; Narvid et al., 2009).

Wallach and colleagues (2010) suggest an approach for solving the affective component of decision making in artificial moral agents. The Learning Intelligent Distribution Agent or LIDA model, is based on machine ethics and cognitive neuroscience, and is inspired by human cognition. The authors distinguish emotions with feelings in this matter, as emotions in LIDA are "…feelings with cognitive content" (p. 466) rather than genuine conscious emotions. The description of emotions and feelings in LIDA are distinct from Damasio's definitions. To acknowledge the difference between feelings and emotions, Damasio explains emotions as an involuntary and fast reaction to an external situation or a mental representation, triggered by a stimulus. Feelings, however, can be prolonged and arises after the emotional reaction as a subjective experience and a *cause* of the emotion (Damasio, 1995). In the LIDA model, a feeling such as pain can be distinguished from the type of pain by nodes on its perceptual memory. For example, if the feeling of pain is from a needlestick or due to an insult, LIDA can distinguish between these different types of feelings. The feelings are categorized according to valence; either positive or negative, and difference in intensity, which make it simple to categorize them. The LIDA model consists of nodes in its perceptual memory that are all unique in their presentation, the valence and intensity of a feeling, is the basis for selecting appropriate behavior. Akin to the human brain, the LIDA model can use association when storing information. For example, if the model perceives negative feelings or actions, the nodes can connect to previous feelings and actions regarding the same feeling, which enables LIDA to form a memory. The stronger the feeling; the probability that LIDA can store it increases (Wallach et al., 2010). Ethical rules such as duties are stored in the semantic memory and are brought up to working memory when needed in the current situation, for example in decision making. The main problem with LIDA, however, is

that association with former events and feelings decays if they are not strong enough, which makes it still not able to compete with the human brain (Cervantes et al., 2016). Nevertheless, in regard to computers and their storage and processing, the computer outperforms the human working memory with its possibility to store short-term memory, or RAM, to a greater extent. Compared to the brain's ability to store memory, the computer's capacity to store long-term memory in the hard drive, is difficult to outperform. The speed of a computer that process large quantities of information, cannot be beaten by the brain as computer architecture often is flat and fast. In contrast, the brain has the ability to prioritize tasks by putting information in a hierarchical manner, but it is more time-consuming (Brooks, Hassabis, Bray, & Shashua, 2012). The core differences between AMAs and humans are the lack of goal-directed behavior and conscious emotions, but as long as one can program rules, values, duties and prejudices, artificial moral agents can still be adequate in moral reasoning (Wallach, 2008). The question is, how much do we need to rely on emotions to be able to make decisions? Is it possible to make decisions without having conscious emotions, such as in AMAs? If referring to what we already know in neuroscience, the amygdala among others parts of the human brain seem to be highly involved in decision making (Shenhav & Greene, 2014). Additionally, deficits of the affective component in moral judgment have shown that it actually *decrease* moral judgment (Damasio et al., 1994; Mendez, 2010; Narvid et al., 2009), which suggest the importance of emotion in moral judgment. Research in cognitive robotics investigates the implementation of trust, episodic memory and specifically; theory of mind in artificial moral agents in relation to artificial moral judgment (Vinanzi, Patacchiola, Chella, & Cangelosi, 2019), as proposed earlier by Wallach and colleugues (2010). By implementing a theory of mind, it would be possible to create agents that understand human behavior, emotions, and intentions, without having conscious emotions themselves. Vinanzi and colleagues (2019) used the Vanderbilt psychological test which aims at investigating the

participant's degree of theory of mind. This test is usually intended for younger children between 3 and 5, but in this experiment, they used a robot with a human-like appearance (humanoid), sitting in front of an instructor. The goal was to investigate whether the humanoid would recognize the instructor's true or false directions of stickers hidden under a left or right paper. The role of the instructor was to play "tricker" by lying, or a helper by telling the humanoid the truth. The humanoid then asked the instructor for suggestions of the sticker locations and decided whether to trust or not trust the instructor's directions. Results indicated that the humanoid, like past studies with 5-year old's that displayed a matured ToM, was able to distinguish when to trust and not trust the instructor with the sticker locations, by also using its episodic memory. This progress in implementing ToM in robots, might be one of the steps towards the implementation of moral cognitive components in non-human agents. However, Wu argues that without the emotional aspects and having a sense of reward and punishment, as humans have with our reward-system, AI technology won't be able to outdo human learning as long as rewards and punishment are not integrated to guide their learning behavior (Wu, 2019).

5.1 The "Hard Problem" for Machine Ethics

"The greater the freedom of a machine, the more it will need moral standards" (Picard, 1997, p. 194). One of the main differences between human beings and machines, is that artificial agents lack conscious emotions and the ability to prioritize tasks and work toward a goal. Humans can solve problems that include insufficient or inaccurate information. If looking at it from a rational AI perspective, these machines have the logical components not influenced by emotions and are still able to make decisions (Wallach, 2008).

A prominent issue with the development of autonomous agents, is the moral dilemma regarding driverless cars. Greene (2016) argues whether how to program vehicles to

make complex decisions. For example, if facing a dilemma where one must navigate a driverless car into a concrete wall, to be able to save pedestrians, in expense of killing the passengers which the car is supposed to protect. In these dilemmas, people often tend to lean towards a utilitarian solution in surveys, namely, to save more people over less such as saving 5 pedestrians over 1 car-passenger. When they get the question whether they would buy a vehicle that would act accordingly, however, the response is usually no. One main point in this dilemma is that participants more often choose to save the passengers instead of the pedestrians if the passengers are family members (Greene, 2016). This highlights the role of emotions in decision making, and how emotions make moral judgment inconsistent. There is also a disagreement issue regarding which moral system which is the most appropriate to implement in machines. Surveys on employees in philosophy faculty have shown that deontology is favored by 26% whereas utilitarianism is favored by 24%. The rest favored virtue ethics or other views, meaning that a compromise might be necessary to be able to implement morality into machines. The problem with this, however, is that as soon as moral agents act according to a specific moral system, people who do not agree with that system will think it is immoral (Bogosian, 2017).

6. Discussion

6.1 A Cognitive Neuroscientific Approach to Artificial Moral Agency

The aim of this thesis is to present the neural correlates of moral cognition and moral judgment from a cognitive neuroscientific perspective, to further discuss the knowledge about the human brain and its importance for the implementation of moral cognition in artificial moral agents. The field of cognitive neuroscience, moral psychology and machine ethics has been extensively investigated through a literature search, with the objective of finding the most relevant knowledge where cognitive neuroscience and machine ethics can

integrate. Moral psychology has enabled these other two fields to integrate, as it covers both ethical dilemmas and deficits in moral cognitive components.

From ethics to evolution, moral theories such as utilitarian and deontological views have been included in many challenging discussions whether which system is the most ethical (Conway et al., 2018; Li et al., 2020; Takamatsu, 2018; Takamatsu & Takai, 2019). With theories such as the dual-process theory, researchers are trying to uncover how the human moral system is functioning. Whether there is a fast or slow moral system, morality is hypothesized to be inherited from our predecessors, as a function that favors prosocial behavior and altruism for humans to survive (Cáceda et al., 2017; Greene, 2014[a]), Reciprocal behavior, specifically, appears to be associated with similar regions as in moral cognition, such as limbic activation but also activation in the MTL, TPJ, PCC and the precuneus, as also seen in decision making and theory of mind (Alcalá-López et al., 2019; Bzdok et al., 2012; Eres et al., 2018; Garrigan et al., 2016).

The moral compass guides human beings into making the right choices, whether the decisions are big or small, happen at the grocery store or in moral dilemmas. Areas such as the vmPFC, dm/dlPFC, TPJ, ACC and amygdala have strong links with cognitive and affective judgment together with theory of mind and cognitive control (Alcalá-López et al., 2019; Bzdok et al., 2012; Eres et al., 2018; Greene et al., 2004). Furthermore, emotions in moral cognition are correlated with activation in the insula, for instance, in evaluation of how to act in harmful or more conventional-based decisions (White et al., 2017).

Researchers have discussed the importance of emotion in decision making, and what happens when emotions are absent during moral judgment. The consequences, are a lack of knowledge about other peoples' emotions and behavior (ToM), but also a decreased ability to empathize and adopting other peoples' emotions (Kiehl et al., 2001; Mendez, 2010; Narvid et al., 2009). Lesion studies have displayed the consequences of the absence of emotions on

moral behavior. Deficits in moral cognition such as moral decision making, empathy and theory of mind, reveals that patients who have lesions to areas such as the vmPFC, ACC and the amygdala, experience less ability to make moral decisions. These patients also have difficulties with understanding other peoples' behavior and emotions, which indicate a major involvement of emotions in our moral cognition (Bechara et al., 1996; Fede et al., 2016; Mendez, 2010; Narvid et al., 2009). Lesions or deficits to areas such as the vmPFC (Bechara et al., 1996; Cameron et al., 2018; Mendez, 2010; Narvid et al., 2009), amygdala (Kiehl et al., 2001) ACC (Kiehl et al., 2001; Mendez, 2010; Narvid et al., 2009), and temporo-parietal junction (Fede et al., 2016) are associated with decreased response to affective stimuli together with bad intentional and unintentional judgment, risky behavior and psychopathic and sociopathic personality traits. Behaviors that all have resulted in immoral actions. This implicates the fundamental aspects of emotion in moral judgment, as both the presence and absence of emotions affect our decisions differently.

In relation to the dual-process theory, Joshua Greene's metaphor of the digital camera on the human brain, can sometimes appear confusing, as the human brain is never completely turned off compared to a camera or a computer (Greene, 2014[a]). As criticism to the dual process theory states, is there simply a fast and slow system that answers to intuitional and rational inputs separately, or are they intertwined (Białek & De Neys, 2017)? In an oppositional perspective, our moral judgment has been associated with distinct neural networks, that differ in activation depending on the emotional influence of the task. Here in which intuitional options in moral dilemmas have shown a stronger connection between the vmPFC and the amygdala compared to pure utilitarian options, suggesting a difference in connections (Shenhav & Greene, 2014). In relation to the dual-process theory, Greene proposes that emotions often appear as an unconscious mechanism of the automatic, "point and shoot" system, meaning that we are not even aware of its influence (Greene, 2014[c]),

which one can argue whether emotions must be *conscious* for us to act accordingly? It may be a "hard problem" for researchers to understand the origins of consciousness, but if most emotions appear non-consciously during moral judgment, why would it be necessary to implement them in machines? As opposed to using the brain as a model for machine intelligence, professor Rodney Brooks in robotics at M.I.T, argues that "Should those machines…" (computers) "…be modelled on the brain, given that our models of the brain are performed on such machines?" (Brooks, Hassabis, Bray, & Shashua, 2012, p. 462). One cannot deny that he has a point. The argument becomes circular, and the discussion takes us back to where we started; do we need all components of human morality to create artificial moral judgment?

Co-existence between human beings and artificial agents are experienced to a great extent nowadays, humanoids, driverless cars and intelligent systems that are implemented in our homes gets more common. Not to mention autonomous agents in industry or in war (Cervantes et al., 2020). The discussion on whether emotions contribute to moral judgment, is more than relevant for the development of artificial moral agency. Is it possible to create a moral agent, despite the lack of conscious emotions? If emotions affect our moral decisions non-consciously, do they actually facilitate our moral judgment? Anderson and Anderson (2007) argue that just as emotion makes us humans "moral beings", emotions also mislead our judgment, making our decisions inconsistent and counteracts the possibility to be restricted to one specific moral system. Research on moral decision making has consistently displayed the incoherence of human decision making (Greene, 2016; Pastötter et al., 2013; Shukla et al., 2019; Trémolière & Bonnefon, 2014), which makes one to question the reliability of the human brain as a model for artificial moral agency. Emotions play a big part in our moral judgment, but we have also seen that humans are prone to be misled by emotions. As presented earlier, positive mood induction prior to moral dilemmas seem to

affect our moral judgment, making it easier to sacrifice humans in moral dilemmas (Pastötter et al., 2013). Additionally, a good mood also increases biases toward specific choices (Shukla et al., 2019). This tendency to change one's mind after the wind, appear to be a common feature of being human. This uncertainty could be due to our ability to empathize, as emotions have a strong influence on how we understand other people (Mendez, 2010; Narvid et al., 2009). From another point of view, humans are not restricted to a pre-programmed moral system as machines, and we have the possibility to make exceptions (Wu, 2019). This is positive in many circumstances, but the reliability of an artificial moral agent would outdo human morality. As we know by now, there is a major disagreement, whether how to act in ethical dilemmas or which moral system to argue as most appropriate (Bogosian, 2017). So how can we circumvent this disagreement and program ethical machines that make appropriate decisions? Big questions, that might have many different answers.

Emotions steer us in different directions depending on the situation (Shukla et al., 2019), intervene when we try to make decisions, but without them, our moral compass would be non-existent and result in impaired moral judgment (Mendez, 2010; Narvid et al., 2009).

To implement morality in artificial agents, it would require an approach to either include emotions in the equation or disavow them. A solution to the lack of emotions in artificial moral agents, would be to implement feelings, as suggested in the LIDA model (Cervantes et al., 2016; Wallach et al., 2010). Wallach and colleagues (2010) also suggest the need for theory of mind to be implemented in the LIDA model for it to function properly with humans in social contexts. By using functions in LIDA such as the understanding of human facial expressions and so on, it would lay the foundation to further implement a theory of mind. In their study, however, that is something that has to be further explored (Wallach et al., 2010). With the ability for AMAs to understand human emotions, behavior, and intentions,

without having them themselves, it would be possible to create an artificial moral agent that at least has feelings considered. On the other hand, should autonomous vehicles have *feelings*? Regarding this matter, it would be necessary to re-evaluate whether vehicles need to understand human emotions through feelings. Would feelings facilitate the decision process when vehicles must decide whether to avoid collision to save its passenger, or to collide with pedestrians? Perhaps, but that also requires a full ethical agent, as mentioned in Moors taxonomy, and for now, only humans can act as full ethical agents (Cervantes et al., 2020; Moor, 2006).

As proposed by current studies, research on theory of mind in robots has evolved and are now implemented in artificial moral agents. With this model suggested by Vinanzi et al. (2019), theory of mind can successfully be implemented in AMAs. Their model of an artificial moral agent integrates both theory of mind and episodic memory. With that in mind, machines can learn from prior experiences, which was only partially developed in LIDA. From another point of view, including episodic memory might increase the risk of prejudices, as past negative experiences can influence and bring biases into future moral judgments. This proposed architecture with the inclusion of ToM suggest that the implementation of affective cognitive components in artificial agents is in prospect. On the other hand, architectures and models that has been tested in moral evalutations are still being restricted to controlled experimental environments, without unexpected situations that can arise. For artificial moral agents to function in harmony with humans, these prototypes have to be further developed and brought out in the external world (Cervantes et al., 2020).

One can argue how important cognitive neuroscience is for the field of artificial intelligence and moral machines. To what extent do we need to understand the neurological basis of the brain, rather than just the cognitive one? LIDA and its architecture is built upon knowledge from machine ethics and human cognition (Wallach et al., 2010), but how does

neuroscience matter for artificial moral agents? The implications of cognitive neuroscience for machine ethics and the development of AMAs are boiled down to the affective components of moral judgment, and how our understanding of the role of emotion affect our moral decisions in different ways. This is where the neuroscientific underpinning of moral behavior matter, because it explains both how rational and affective decision making are part of our moral judgment. From another point of view, moral psychology would need a common ground to meet before we can integrate moral systems in AMAs. If we do not have a clear answer to what is immoral or not, how can we implement morality in machines? We have seen that there is huge disagreement in which moral system to view as the most appropriate, both among professionals in philosophical faculty as well as with lay people and their reactions to the Trolley and Footbridge dilemmas. There are arguments whether a utilitarian or a deontological approach is "more" ethical, due to disagreement with the emotional aspects of moral judgment. In some sense, utilitarian options appear to be most ethical when weighing cost-benefit, whereas some argue that these options are more correlated to impaired empathy rather than an ethical behavior (Takamatsu, 2018; Takamatsu & Takai, 2019). There is a risk that if we program moral machines based on specific moral theories, these machines are not going to be able to act ethical because there is always going to be disagreement which theory and which choice that is the most appropriate (Bogosian, 2017). This is where cognitive neuroscience matters for artificial moral agency. Implementing moral systems such as utilitarian or deontological views may be a necessity for AMAs, but we also know that human morality derive from neural and cognitive structures and functions in the human brain, as deficits regarding these structures display unethical behavior. Functions such as decision making, cognitive control, empathy, ToM and episodic memory are equally important for understanding moral behavior as the philosophical underpinnings and theories in normative

ethics. This, emphasizes the importance of cognitive neuroscientific resarch for machine ethics and machine intelligence.

How we intertwine cognitive neuroscience, moral psychology and artificial intelligence in the future is of substantial importance, because the development of artficial intelligent agents are rather going to increase than the opposite. However, by integrating these fields, it comes with several limitations as well. No matter how important the brain is for understanding morality, there are still problems that we are facing when it comes to the implementation of emotions, choosing the correct moral system to implement into machines, how to program autonomous vehicles to act ethically and so forth. These questions underscore the importance of understanding the philosophical underpinnings as well as the human brain and the neural correlates of moral judgment.

Perhaps, future research will be able to implement a dual-process morality into AMAs, having both the rational and affective components of moral cognition in mind, as well as the ethical underpinnings. However, an implementation of a dual-process morality can also bring dilemmas regarding the inconsistency during emotion-influenced judgment, which would result in AMAs to be just as indecisive as humans. This questions the implementation of the affective component of moral judgment. Emotions are of substantial importance for moral behavior, but the implications of this implementation will perhaps be to grand, and bring further, more complex dilemmas.

7. Conclusion

The neural correlates of moral cognition and moral judgment; how we make complex decisions, interpret and understand other people's intentions and emotions, empathize with other people, as well as acting altruistic, are key components of acting moral. The neural structures behind morality are interconnected to a great extent, as they are engaged in both rational and intuitional moral judgment. These moral functions and behaviors, are unique to human beings, but as technology and engineering advances, so does artifical intelligence and the implementation of morality into machines. The development of artificial moral agents and the integration of cognitive neuroscience might be problematic in some ways, especially in regards to concious emotions. From studies that have been introduced in this thesis, results display that emotions play a major role in our moral judgment. Whether or not we will be able to implement emotions, feelings, or theory of mind in moral machines, we know that the absence of emotions in humans have resulted in several immoral behaviors. In the contrary, emotions seem to blur and bias our judgment. This indicates the importance for future research of the affective components of moral judgment as well as for machine ethics, both in terms of its influence on decision making but also emotions and its significance for acting ethically. The answer could be to meet half-way, with moral agents that have an integrated ToM; can learn from past experiences with episodic memory, and agents that are able to select appropriate actions depending on moral systems and human interaction. Until then, cogntive neuroscience, machine ethics and engineering will have to co-operate. As the human brain appears to be the best model we have for developing artificial moral behavior, humans seem to be the only being at the moment that are capable of acting like a full ethical agent. Again, this underscores the importance of understanding the neural correlates of moral judgment for the development of moral machines.

References

Abe, N., Greene, J. D., & Kiehl, K. A. (2018). Reduced engagement of the anterior cingulate cortex in the dishonest decision-making of incarcerated psychopaths. *Social Cognitive and Affective Neuroscience*, *13*(8), 797–807. doi: 10.1093/scan/nsy050

Alcalá-López, D., Vogeley, K., Binkofski, F., & Bzdok, D. (2019). Building blocks of social cognition: Mirror, mentalize, share? *Cortex*, *118*, 4–18. doi:10.1016/j.cortex.2018.05.006

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. doi: 10.1080/09528130050111428

Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent. *Ai Magazine*, *28*, 15–26.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. R. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex*, *6*(2), 215. doi: 10.1093/cercor/6.2.215

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167.

Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. *Minds & Machines*, *27*(4), 591–608. doi: 10.1007/s11023-017-9448-z

Brooks, R., Hassabis, D., Bray, D., & Shashua, A. (2012). Turing centenary: Is the brain a good model for machine intelligence? *Nature*, *482*(7386), 462–463. doi: 10.1038/482462a

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A., Langner, R., & Eickhoff, S.

(2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function*, *217*(4), 783–796. doi: 10.1007/s00429-012-0380-y

Cáceda, R., Prendes-Alvarez, S., Hsu, J.-J., Tripathi, S. P., Kilts, C. D., & James, G. A. (2017). The neural correlates of reciprocity are sensitive to prior experience of reciprocity. *Behavioural Brain Research*, *332*, 136–144. doi: 10.1016/j.bbr.2017.05.030

Cameron, C. D., Reber, J., Spring, V. L., & Tranel, D. (2018). Damage to the ventromedial prefrontal cortex is associated with impairments in both spontaneous and deliberative moral judgments. *Neuropsychologia*, *111*, 261–268. doi: 10.1016/j.neuropsychologia.2018.01.038

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501–532. doi: 10.1007/s11948-019-00151-x

Cervantes, J.-A, Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation*, *8*(2), 278–296. doi: 10.1007/s12559-015-9362-8

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*. doi: 10.1016/j.cognition.2018.04.018

Damasio, A. R. (1995). *Descartes' error : emotion, reason, and the human brain.* Avon Books.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damasio, A. R. (1994). The

return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science, 264*(5162), 1102–1105. doi: 10.1126/science.8178168

Decety, J., & Wheatley, T. (2015). *The Moral Brain: A Multidisciplinary Perspective*. Cambridge, Massachusetts: The MIT Press.

Eres, R. L., W. R., & Molenberghs, P. (2018). Common and distinct neural networks involved in fMRI studies investigating morality: an ALE meta-analysis. *Social Neuroscience, 13*(4), 384–398. doi: 10.1080/17470919.2017.1357657

Fede, S. J., Borg, J. S., Nyalakanti, P. K., Harenski, C. L., Cope, L. M., Sinnott-Armstrong, W., … Kiehl, K. A. (2016). Distinct neuronal patterns of positive and negative moral processing in psychopathy. *Cognitive, Affective, & Behavioral Neuroscience, 16*(6), 1074–1085. doi: 10.3758/s13415-016-0454-z

Fujino, J., Tei, S., Itahashi, T., Aoki, Y. Y., Ohta, H., Kubota, M., … Nakamura, M. (2020). Role of the right temporoparietal junction in intergroup bias in trust decisions. *Human Brain Mapping, 41*(6), 1677–1688. doi: 10.1002/hbm.24903

Garrigan, B., Adlam, A. L. R., & Langdon, P. E. (2016). The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements. *Brain and Cognition, 108*, 88–97. doi: 10.1016/j.bandc.2016.07.007

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*. doi: 10.1016/j.neuron.2004.09.027

Greene, J. D. (2014[a]). *Moral tribes: Emotion, reason, and the gap between us and them.* Atlantic Books.

Greene, J. D. (2014[b]). The cognitive neuroscience of moral judgment and decision-making. *The Cognitive Neurosciences*, *6*.

Greene, J. D. (2014[c]). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*(4), 695–726. doi: 10.1086/675875

Greene, J. D. (2015). The rise of moral cognition. *Cognition*, *135*, 39–42. doi: 10.1016/j.cognition.2014.11.018

Greene, J. D. (2016). Our driverless dilemma. *Science*, *352*(6393). doi: 10.1126/science.aaf954

Haidt, J. (2008). Morality. *Perspectives on Psychological Science*, *3*(1), 65–72. doi: 10.1111/j.1745-6916.2008.00063.x

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality & Social Psychology*, *79*(6), 995–1006. doi: 10.1037/0022-3514.79.6.995

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2017). Individual differences in decision making depend on cognitive abilities, monitoring and control. *Journal of Behavioral Decision Making*, *30*(2), 209–223. doi: 10.1002/bdm.1939

Kant, I. (2011). In M. Gregor, & J. Timmermann (Eds. & Trans.), *Immanuel Kant: Groundwork of the Metaphysics of Morals: A German–English Edition*. Cambridge, UK: Cambridge University Press.

Kiehl, K. A., Smith, A. M., Hare, R. D., Mendrek, A., Forster, B. B., Brink, J., & Liddle, P. F. (2001). Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. *Biological Psychiatry*, *50*(9), 677–684. doi: 10.1016/S0006-3223(01)01222-7

Li, S., Ding, D., Lai, J., Zhang, X., Wu, Z., & Liu, C. (2020). The characteristics of moral judgment of psychopaths: The mediating effect of the deontological tendency. *Psychology Research and Behavior Management, 13,* 257–266. doi: 10.2147/PRBM.S226722

Mendez, M. F. (2010). The unique predisposition to criminal violations in frontotemporal dementia. *The Journal of the American Academy of Psychiatry and the Law, 38*(3), 318–323.

Mill, J. S., & Corporation, E. (2011). *Utilitarianism*. Luton: Andrews UK.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18–21. doi: 10.1109/MIS.2006.80

Narvid, J., Gorno-Tempini, M. L., Slavotinek, A., DeArmond, S. J., Cha, Y. H., Miller, B. L., & Rankin, K. (2009). Of brain and bone: The unusual case of Dr. A. *Neurocase*, *15*(3), 190–205. doi: 10.1080/13554790802632967

Norman, D. A., & Shallice, T. (1980). Attention to action: Willed and automatic control of behavior technical report no. 8006.

Pastötter, B., Gleixner, S., Neuhauser, T., & Bäuml, K.-H. T. (2013). To push or not to push? Affective influences on moral judgment depend on decision frame. *Cognition*, *126*(3), 373–377. doi: 10.1016/j.cognition.2012.11.003

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality & Social Psychology, 64*(3), 467–478. doi: 10.1037/0022-3514.64.3.467

Picard, R. W. (1997). *Affective computing.* MIT Press.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the

amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(13), 4741–4749. doi: 10.1523/JNEUROSCI.3390-13.2014

Shukla, M., Rasmussen, E. C., & Nestor, P. G. (2019). Emotion and decision-making: Induced mood influences IGT scores and deck selection strategies. *Journal of Clinical & Experimental Neuropsychology*, *41*(4), 341. doi: 10.1080/13803395.2018.1562049

Takamatsu, R. (2018). Turning off the empathy switch: Lower empathic concern for the victim leads to utilitarian choices of action. *PLoS ONE*, *13*(9). doi: 10.1371/journal.pone.0203826

Takamatsu, R., & Takai, J. (2019). With or without empathy: Primary psychopathy and difficulty in identifying feelings predict utilitarian judgment in sacrificial dilemmas. *Ethics and Behavior*, *29*(1), 71–85. doi: 10.1080/10508422.2017.1367684

Tonkens, R. (2009). A challenge for machine ethics. *Minds & Machines*, *19*(3), 421–438. doi: 10.1007/s11023-009-9159-1

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality & Social Psychology Bulletin*, *40*(7), 923–930. doi: 10.1177/0146167214530436

Turing, A. M. (1950). *Computing Machinery and Intelligence.*

Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771). doi: 10.1098/rstb.2018.0032

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & Society*, *22*(4), 463.

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, *2*(3), 454–485. doi: 10.1111/j.1756-8765.2010.01095.x

White, S. F., Zhao, H., Leong, K. K., Smetana, J. G., Nucci, L. P., & Blair, R. J. R. (2017). Neural correlates of conventional and harm/welfare-based moral decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(6), 1114–1128. doi: 10.3758/s13415-017-0536-6

Wu, Y. (2019). Research on the development of integration of neuroscience and artificial intelligence. *IOP Conference Series: Earth and Environmental Science*, *384*, 12007. doi: 10.1088/1755-1315/384/1/012007