Moral Psychology and Artificial Agents (Part 1):

Ontologically Categorizing Bio-Cultural Humans

Michael Laakasuo, Jukka Sundvall, Anton Berg, Marianna Drosinou, Volo Herzon, Anton Kunnari, Mika Koverola, Marko Repo, Teemu Saikkonen and Jussi Palomäki.

This is an article in press, full reference:

Laakasuo et al. (in press). Moral Psychology and Artificial Agents (Part 1): Ontologically Categorizing Bio-Cultural Humans. *Machine Law, Ethics and Morality in the Age of Artificial Intelligence*. Steven Thompson (ed). New York: Igi Global.

INTRODUCTION

We are surrounded by autonomous artificial agents, many of which pose previously unseen moral challenges through their actions and the consequences of these actions. Artificial intelligences (AIs) are already used in medical diagnostics (Mangasarian, Setiono & Wolberg, 1990; Fjell et al., 2008), financial credit evaluation and approval (Poon, 2007), traffic and transportation (Li et al., 2016), and numerous military applications (Springer, 2013). There are visions of life-saving robots guarding our beaches (Guillette, 2019), nanobots cleaning our polluted oceans (Singh & Naveen, 2014), and even robot prostitution (Levy, 2007). But what if an AI recommends medication against a patient's will, or denies a life-changing loan to a family in dire need? What if an autonomous vehicle carrying children decided to divert into a ditch to avoid hitting elderly pedestrians? Should one be allowed to "cheat" on their spouse with a robot, or better yet, marry one? What if military drones could decide on their own where to unload their weapons? These questions of responsibility for autonomous decision-making and behavior that belong to "robot morality" have exceptional societal relevance, not only in terms of law and regulation, but also for our future as a species.

Despite our modern surroundings, however, human cognition is shaped by evolution. What this means in practice, is that humans have fundamental intuitive, automatic, and non-conscious processes constantly operating in the background. Such processes organize our perceptions, thoughts, and reactions to the world outside our minds. Here we argue that robots and AIs should be evaluated not only through the lens of analytical moral philosophy, but also using the tools of experimental moral psychology and evolutionary processes.

Understanding what people in general (not only academics) think of moral and legal issues regarding AI helps anticipate them and could (or should) inform technological development and legislation.

In this chapter, we will provide a theoretical background for this current discussion taking place in a variety of different journals. Of late, an increasing number of empirical studies have focused on moral issues related to AIs and opined on transhumanistic concerns for the future of our species. Research has begun uncovering previously unseen moral cognitive phenomena, such as new types of cognitive biases in human-AI interaction, thereby challenging existing theoretical frameworks in cognitive science. In order to describe these challenges, we need to provide the reader with an extensive review of evolutionary, moral, and cognitive sciences of concepts and categories. In another chapter of this book, we review empirical research in the emerging field of moral psychology of robotics and transhumanism.

Within the last few years, empirical studies have focused on topics such as cross-cultural and cross-geographical differences in moral preferences concerning self-driving cars (Awad et al., 2018), the moral psychology of sex-robots and nursing robots (Koverola et al., 2020; Laakasuo, under revisions), attitudes towards military drones making autonomous decisions as compared with people making the same decisions (Malle et al., 2019), and attitudes toward brain implant technology (Castelo et al., 2019) or mind upload technology (Laakasuo et al., 2018). All of these themes were predicted by several prescient philosophically oriented scholars (see Allhoff et al., 2011 for a review), however we have now come to a point where the topics are being investigated from new moral psychological angles.

Novel findings from these studies show, for example, that people appreciate hypothetical decisions made by robots more if the robots are perceived to have human-like

minds (Bigman & Gray, 2018; Malle et al., 2019). This is peculiar because a decision to save somebody, for example, from drowning, is the same decision superficially - from a third-person perspective - whether it is made by a human lifeguard, a trained rescue dog, or an autonomous life-saving robot. One proposed explanation draws from a family of cognitive phenomena known as *mind perception mechanisms* in the context of robotic interactions (Bigman & Gray, 2018): robots are perceived as having less cognitive or emotional capabilities, and thus their decisions are more suspect. However, in our own studies we have observed that mind perception does not fully explain our aversion towards robots making moral decisions. Instead, we have found that the *type* of moral decision matters in terms of mind perception (See section 3 and Part 2; Laakasuo et al., under revisions). Current theories also fail to explain recent findings in research on transhumanistic themes and whether the problems observed in the context of robots extend to androids and cyborgs as well. For example, why is it that the aversion towards "mind upload" technology (making a digital copy of the brain into a computer) and silicon-based brain implant technology ("chips in the brain") seems to be driven by sexual disgust sensitivity (Laakasuo et al., 2018; Koverola et al., 2020)? (see Part 2).

Technologies that merge the human body and the human mind/brain (or mimic them in some ways) with different forms of information processing machines are usually defined as transhuman technologies. They are called transhuman because they are perceived as something that alters the fundamentals of the human nature and allows humans to take control of their own evolutionary processes. However, technologies like uploading one's mind into a computer or substituting parts of one's brain with silicon-based chips, also cause categorical confusions. When does a human become a machine and when does a machine become human? How do we feel, morally, about things that deeply challenge the fundamentals of being human? In this chapter, we dive deep into the cognitive science of

categories and categorization, and aim to give the reader a theoretical background that helps them come to terms with recent empirical findings that we will introduce in Part 2.

In section 2 here, we briefly introduce the basics of evolutionary psychology as it offers a coherent view of human cognitive capacities and their coevolution with tool creation, and it is connected to both the field of moral psychology and the question of how people perceive non-human agents. In section 3, we provide an overview of moral psychology and its current most important theoretical models and questions. Section 4 focuses on the innate human tendency for categorical thinking and category formation. This theme is explored further in section 5, where we consider how contemporary robots and AIs do not easily fit within existing mental categories, and how this can lead to systematically biased judgment. We conclude Part 1 with a quick summary of the themes covered.

In Part 2, we briefly cover transhumanism as a philosophy, and consider how transhumanistic technologies cause further categorical confusions that lead to previously unseen moral psychological conundrums. We bring the previous themes together in a review of moral psychological literature pertaining to both transhumanistic technologies and intelligent artificial agents. At the end of Part 2, we propose some new directions for the study of human moral cognition relating to robots and AIs, and summarize the implications of the themes of Parts 1 and 2 for cognitive science and related fields.

2. EVOLUTIONARY PSYCHOLOGY – HUMAN SOCIAL COGNITION AND OTHER MINDS

Humans have evolved under a myriad of different conditions and ecological niches during the last six million years. The most significant evolutionary changes in human cognitive capacities took place in the Pleistocene epoch, between approximately 1.8 million and 10,000 years ago (see e.g. Lee & Wolpoff, 2003; Shultz et al., 2012). During this time, anatomically

modern humans developed symbol use, cave painting techniques, and language (Lewis-Williams, 2002). The cranial capacity of *Homo sapiens* has not significantly changed in about 200,000 years (Dunbar, 2014).

Evolutionary psychology (EP) argues that modern *Homo sapiens* are, by and large, the same all around the planet (Pinker, 1997; 2002). EP seeks to find what is universal and common in all humans by studying their cognitive and behavioral similarities across cultures. EP aims to provide evolutionary explanations to these similarities, based on selection pressures and challenges that humans went through during the Pleistocene era (Tooby & Cosmides, 2005). Many cultural phenomena can be thus explained as by-products of our evolutionary background (Tooby & Cosmides, 1992). EP thus focuses on human cognitive universals. To paraphrase anthropologist Roger Keesing (see Keesing & Strathern, 1997), humans are equally rational wherever they scratch their behinds; only the ways in which this rationality is expressed are different on the surface.

Although humans have evolved in various ecological environments with various challenges to survival and reproduction, certain challenges have remained constant. For example, there has always been rain, sunshine, hunger, pathogens, plant toxins, predators, competition for mates, violence, child rearing, hunting, food gathering, group-coordination, communication, and tool use and creation (Tooby & Cosmides, 2005). All of these challenges create selection pressures, and, thusly, humans have evolved specific cognitive mechanisms to deal with those challenges. EP argues that humans have evolved cognitive adaptations for specific functions (i.e. *modules* of cognition) such as mate selection (interest towards the preferred sex), avoiding sources of pathogens (the emotion of disgust), and avoiding predators or other dangers (the emotion of fear). Nonetheless, probably the most relevant or

interesting cognitive aspects from an evolutionary perspective for humans, in the context of this chapter, are socio-cognitive.

Group living is one of the fundamental constants in human evolutionary history. There has always been some element of resource sharing with one's kin, extended kin, and friends (Dunbar, 2014). Humans are born into this world probably more helpless than any other mammals (Baumeister & Leary, 1995); the simple act of walking takes humans about a year to learn, whereas for elephants, horses, and antelopes it takes some hours at most. The human socio-evolutionary environment is complex and it takes environmental fine-tuning of cognitive adaptations for several years before the individual can survive on its own (Pinker, 1997). As an example, the human capacity for language is biological – it grows out of people, like breasts and beards do (requiring nutritional resources from the environment to fully develop) – but the language humans learn as their primary one is obviously controlled by the environment (Pinker, 1994).

Similarly, humans have evolved as organisms that have several socio-cognitive instincts to solve problems of group and tribal living. In this context, EP uses the term *instinct* in a very specific way: automatic cognitive processing that happens fast, and largely outside of consciousness, and that is only felt in the mind as feelings. In this sense, our instincts are like the feeling of hunger, which is an output of a complex psycho-physiological machinery keeping track of multiple variables. On the physiological level, the ”state” of hunger is a complex interaction pattern between different kinds of peptides, hormones, neurotransmitters, and neural impulses taking place between the digestive system and the central nervous system.

A similar logic applies to human social instincts. Feelings of obligation (attachment, care, etc.) towards one's offspring, the feeling of guilt when one has transgressed on

another's well-being, or the feeling of embarrassment when one has committed a norm violation in public, are similar to hunger. Social instincts have functions, but we may not be acutely aware of the logic behind them (Keltner, Haidt, & Shiota, 2006). The function of hunger is to guide the organism towards calorie consumption. Calories enable it to survive until the opportune moment for mating presents itself, so that the genes responsible for that instinct can make copies of themselves[1]. Similarly, the sense of obligation towards one's children makes it more likely that younger generations are raised until the age of reproduction. The function of guilt is to motivate the individual to signal remorse and make amends so that they are not ostracized by the larger community from vital resources relevant for survival and mating opportunities (Breggin, 2015). Similarly, embarrassment functions as a signal to others in one's community that a norm was accidentally violated and that one is not a threat (Feinberg et al., 2011), thus preventing potentially violent situations from escalating, avoiding potential costly physical damage.

These and other socio-cognitive adaptations mostly regulate the social environment where the individual is embedded, and ultimately serve survival (Francis, 1990). In contrast, there is a cognitive deficiency known as prosopagnosia (Grüter et al., 2008); that is, impaired facial recognition. People with prosopagnosia can survive and reproduce, but ultimately there are survival-related advantages in being able to tell people apart based on their faces.

For the area of moral psychology (see section 3), one of the central socio-cognitive mechanisms that humans have more than other great apes is the capacity to think about other minds (Dunbar, 2014). Without going into technicalities of how this capacity has developed (Dunbar, 2014; see also Tsoukalas, 2018, for a more fundamental evolutionary hypothesis), it allowed proto-human apes to keep tabs on the social reputations of their partners and stay

1. Naturally, we acknowledge that genes themselves do not think, feel, or act in any way. However, this form of expresson is a convenient short-hand, which only means that some gene variants in the gene pool are, on average, in comparison to all the other variants, more efficient in increasing in their relative frequency.

vigilant about not being exploited in the social exchange of favors (i.e., mostly grooming behavior). The ability to reason about other minds may also have been more immediately beneficial for the survival of one's offspring, as some recent research suggests a link between a mother's capacity for thinking about other minds and their sensitivity to their child, which in turn is predictive of developmental cognitive capacities (Licata et al., 2016; Rigby et al., 2016; Zeytinoglu et al., 2018). Whatever the evolutionary origins or earliest functions of the capacity to think of other minds, this capacity is relevant in many areas of human social interaction.

The capacity to assign mental states and emotions to others and think about them is commonly referred to as *theory of mind* (ToM; see, e.g., Saxe & Baron-Cohen, 2006). People who lack this specific ability, or who have some deviation in this ability from the population average, are often diagnosed with an autism spectrum disorder and have higher tendencies of treating other living beings as objects or in a more objectifying manner (Saxe & Baron-Cohen, 2006). However, ToM research largely focuses on how (and whether) people ascribe specific mental *contents* to others. We argue, in line with Gray, Young, and Waytz (2012) that ascribing any kind of mind that could even have those mental contents - *mind perception* - is more fundamental, even if similar cognitive processes are behind both capacities. That is, one has to be able to view someone or something as having or not having a mind (i.e., differentiate between agents and non-agents) before being able to (correctly or incorrectly) ascribe contents to that mind. Any of the human socio-cognitive skills we have mentioned here would not be able to develop without the ability to perceive others as thinking and feeling agents in the world. Thus, mind perception is a central evolved cognitive feature, with relevance to evolutionary and moral psychology in general, and the moral psychology of robotics more specifically: artificial agents are very different from the kinds of minds people

have usually perceived until now and in our evolutionary history. We return to these themes throughout the rest of this chapter.

3. WHAT IS MORAL PSYCHOLOGY?

Moral psychology studies cognitive processes and structures in humans (and other animals) that are related to decisions, judgments, negotiations, and actions regarding what is considered to be right and wrong in a given context (Voyer & Tarantola, 2017). Moral philosophers use different types of *a priori* methods – like philosophical conceptual analysis and philosophical intuition – to test propositions that follow from moral theories for logical coherence with intuitions (top-down). Or abduct a proposition from a set of moral intuitions (bottom-up) and then test the logical coherency with moral propositions from other theories. They do this to dissect and analyze different types of social situations, and commonly reach conclusions or recommendations in their analyses regarding what should be done about a specific topic. Unlike moral philosophy, moral psychology usually does not take explicit positions regarding what is objectively, or by normative standards, right/wrong.

Moral psychologists mostly aim to observe and measure the expression of ordinary people, children, and professional experts in their judgment, reasoning, and deduction in situations which are commonly related to the well-being of other sentient beings[2,3] . Until quite recently, moral psychology was mostly focusing on how humans treat other humans, but as of late, animal rights-related questions have also received attention (e.g., Loughnan et al, 2010). Part 2 of this chapter introduces the recent expansion of moral psychology to new

2. Naturally, harming or benefitting a person is not the only morally relevant action and does not cover everything that can be placed under the umbrella of "morality."

3. To put it shortly, moral psychology is mostly a desciptive science: it describes how morality happens in the the observable universe, but does not "take sides." In this sense, moral psychology, or the field of moral cognitition more broadly, is similar to the study of history, where the historian might describe the attrocities of genocides accurately and carefully, without actually supporting such horror.

subjects: robots (Awad et al., 2018) and transhumanism (Castelo et al., 2019; Laakasuo et al., 2018).

There are two traditions in moral psychology (Haidt, 2007; 2010). The First Wave of moral psychology – labelled *Kohlbergian moral psychology* – studies how morality develops in adults and in children. In this tradition, the research was mostly conducted with deep probing interviews and careful scoring of the level of abstraction the participants used in their speech in order to justify certain moral actions. It was assumed that morality develops in a step-by-step fashion, beginning from a concrete fear of punishment, and then advancing towards abstract universal moral principles applied consistently in various situations (Piagetian and Kohlbergian tradition, Helkama, 2009).

The focus in this book chapter, as well as in our research, is more generally anchored in the Second Wave of moral psychology. In many ways, the Second Wave tradition is an off-shoot, or a parallel development of EP (Haidt, 2007; 2010). In most research in this area, EP is accepted as the background theory more or less explicitly, even if the topics and themes studied in moral psychology are not immediately, obviously related to the core topics usually studied in EP (e.g., mating-related cognition). However, moral psychology in many ways continued the themes related to EP by expanding them, including themes like altruism, cooperation, intergroup helping (Laakasuo et al., 2018), or condemnation of out-group behaviors (Cohen-Chen et al., 2014).

Research conducted in this tradition often utilizes moral dilemmas or *vignettes* (specifically crafted stories or situational descriptions) that juxtapose different types of moral intuitions (Cushman & Greene, 2012). In this tradition, it is also common to use measurement tools and theories developed by personality psychologists, behavioral economists, evolutionary psychologists, and neuroscientists. Second Wave moral psychology is

understood as "hard quantitative" science as it employs statistics much more than qualitative methods. Modern moral psychology is also theoretically and methodologically closely-knit with research on decision-making and emotions. It is also very multidisciplinary. For example, clinically oriented moral psychologists might be interested in the emotional lives of psychopaths and narcissists, and how they solve moral dilemmas compared with "normal people" (see, e.g., Kahane et al., 2015; Tassy et al., 2013). Furthermore, theologically oriented researchers might be interested in how people's free will beliefs predict altruism (Sinnott-Armstrong, 2014). In other words, moral psychology is a broad area within the cognitive sciences, and the topics it studies can be approached from different angles, bringing together clinicians, neuroscientists, philosophers, legal scholars (Mikhail, 2007), and anthropologists (see also Zalta, 2020, for further discussion on definitions of moral psychology).

3.1 Central Models of Moral Cognition.

There are a few models which form large partitions of the core of the Second Wave Moral Psychology. It is essential to understand these basic concepts, so that the themes covered in this book chapter can be understood in their proper context. We will briefly summarize some of them here, however, by no means is this list exhaustive[4].

3.1.1 The Theory of Dyadic Morality (TDM). The TDM has been developed by Kurt Gray and his colleagues in several publications (Gray et al, 2007; 2012; Schein & Gray, 2018). According to the TDM, *mind perception* is the central cognitive mechanism enabling moral cognition (see Section 2). Mind perception refers to the automatic tendency of healthy humans to perceive mental capacities in other living beings, especially in people. Humans

4 With moral cognition, we mean that multidisciplinary area of moral psychology that focuses on defining, explicating and studying the information processing aspects of moral judgments and decisions.

project onto other living beings the same mental capabilities that they themselves have, such as: a) the capacity to experience suffering; b) the capability to feel motivated, and c) the proclivity to act in goal-oriented, sensible, or meaningful ways.

The TDM argues that humans perceive and interpret social situations as morally meaningful if four requirements are met: 1) there is an intentional *moral agent* who 2) causes harm to 3) another being (i.e., a *moral patient*) who 4) has the capacity to experience suffering. Thus, the (fuzzy) template of a morally relevant event always contains an agent, a patient, and a harm. Importantly, this dyadic template allows for a kind of "filling-in." The TDM claims that perceiving some elements of the dyadic interaction between the perpetrator and the victim can lead one to interpreting that the other elements were present as well. This is the TDM's explanation for why people sometimes condemn seemingly harmless acts, and how people may come to infer the presence of an agent causing harm from the presence of a patient experiencing harm, or vice versa.

Consider the so-called "moral dumbfounding" effect. A famous example from Haidt and Hersh (2001) concerns a fictional scenario where two siblings have consensual, non-reproductive sex. Many people unsurprisingly find this wrong, but cannot seem to articulate why: they are dumbfounded, as the act has been carefully described as having no harm. Schein & Gray (2018), in defense of the TDM's stance that moral condemnation implies harm, argue that dumbfounding is an artifact of psychological experiments. That is, an experimenter may tell someone that a hypothetical case of incest was *objectively* harmless, but those who condemn the act do not believe the experimenters: harm is *subjective* to the person judging. Thus, if it turned out that people who find "harmless incest" a believe scenario still condemned it, this would be evidence against the TDM.

The TDM seeks to maximize explanatory power while remaining as simple as possible. The central point is that perceiving moral violations depends on the ability to perceive mental capacities for agency and experience in others, and a flexible cognitive schema for what "harm" looks like. The model seeks to explain what happens in moral condemnation (or moral judgment processes); the model makes no claims as to what people may consider morally praise-worthy. Furthermore, the model is explicitly constructionist and does not tie individual differences in morality to any modular EP framework (Schein & Gray, 2018). That is, unlike, e.g., the Moral Foundations Theory (see below), the TDM claims that morality stems from one basic cognitive template instead of several distinct, innate moral concerns. The TDM grants that cultural, political, and individual differences in what people find harmful, and thus morally condemnable, are possible, but it always ties these differences back to perceptions of agents, patients, and harm.

The TDM has been criticized for being at times circular and its assumptions of necessary elements to a moral violation not being necessary or sufficient (see e.g. Alicke, 2012 and Monroe et al., 2012, for critiques; see also Schein & Gray, 2018, for a response from the developers of the TDM). Critics have argued, among other things, that not all intentional harmful actions are considered immoral (e.g., soldiers who kill their enemies in a war), and that not all harmful actions considered immoral are intentional (e.g., gross negligence). We will not delve deeper into the debate here: our intention is simply to introduce a model that is relevant to discussion in Part 2 of this article.

3.1.2 Moral Foundations Theory (MFT). The Moral Foundations Theory (MFT) posits that human morality consists of five different "foundations" or "taste receptors," each having their corresponding domain of functioning (Graham et al. 2011, 2013; Haidt, 2012; Haidt et al., 2009). According to the MFT, when people evaluate the acceptability of each other's

actions, they pay attention to the following issues: 1) Did the action cause harm? 2) Was the action fair? 3) Was the action respectful towards authorities? 4) Was the action loyal towards the agent's in-group? 5) Was the action "pure" or did it violate holy values?[5] We will refer to these five key issues, respectively, as the foundations of harm, fairness, authority, loyalty, and (moralistic) purity (their labeling, but not content, has changed over time).

Contrary to the TDM, in the MFT the moral condemnation of an action does not presuppose that the action must harm some conscious entity, or fit a catch-all cognitive template. The MFT also assumes that humans differ in their tendencies to favour the five types of moral domains. According to Haidt (2012), liberally oriented people mainly favor the first two foundations: harm and fairness. Conservatives, on the other hand, seem to care about all five foundations equally; but they especially care much more about the (moralistic) purity foundation than liberals. Conservatives, for example, condemn the burning of their country's flag and oppose cannabis use more than liberals. Liberals, on the other hand, condemn income inequality and the intentional widening of the income gap.

The MFT is explicitly modular and thus more closely tied to the EP framework (Haidt, 2012). Each of the foundations is considered a separate cognitive module with specific types of inputs (events in the world, perceived as a violation of a specific foundation) and outputs (foundation-specific moral judgment). To return to the example of "harmless incest," the MFT would explain condemnations of this event as stemming from the event being a violation of the purity foundation. Furthermore, the reason for differences in judgments of acts that violate this foundation is simply that different groups of people develop to emphasize the purity foundation differently.

5 We understand that this list is not exhaustive from a philosophical perspective; i.e., it lacks many of the classical themes that moral philosophy is based upon, such as respect for individual liberties, rights, and respect for the dignity of others. However, this is the theory in its current formulation and it is based on empirical investigations.

As can be expected, the MFT has been criticized by proponents of the TDM, who claim the foundations can be reduced to concerns about harm (Schein & Gray, 2018). There are also alternative models with different foundations argued to have better evolutionary psychological footing than those of the MFT (Curry et al., 2019). Again, we will not delve deeper into the debate, as we simply wish to introduce an influential model that will be relevant for later discussion.

3.2 Do Moral Emotions Guide Moral Decision-Making?

The science of moral cognition is a hotly debated area. Simplifying slightly, one of the areas of debate regards the question whether moral decision-making is mostly motivated by emotions or by reason (Greene, 2013). Haidt (2001, 2007) has previously suggested that moral judgments are mostly based on emotions and intuitions: we condemn actions because they evoke negative emotions, such as disgust or anger. While this emotion-driven view has been impactful, some researchers are still in favor of moral reasoning as the most important component (e.g. Mikhail, 2007). Recently, McAuliffe (2019) argued that the evidence for the causal role of emotions in moral judgment is weaker than assumed, and that existing research on emotions and morality sometimes actually supports rationalist theories of moral thought.

The role of emotions in moral judgment is highlighted in studies of utilitarian moral judgment. From a utilitarian (or consequentialist) perspective, killing can be justified if by killing one, several others are saved (see, e.g., Greene, 2007; 2013). According to the utilitarian morals (Mill,1861; Bentham, 1816; Sidgwick, 1874), it is imperative to try to maximize the amount of "good" regardless of the specific act. Utilitarian morality is commonly juxtaposed with deontological morality, which focuses on absolute rules, principles, and obligations to perform or omit a certain action; regardless of the situation (Greene, 2013). Deontological morality has been argued to be more reliant on emotions

(either implicitly or explicitly) than utilitarian "moral calculus." The juxtapositioning of these two moral standings supposes that humans are primarily *either* deontological *or* utilitarian in their judgments (but not necessarily in their stated moral philosophies, if they have any). Previous neuroimaging research implied that deontological moral evaluations were made faster than utilitarian ones (Greene, 2013). Indeed, utilitarian evaluations correlated with "higher cognition" brain activation (i.e., activation in areas related to working memory, rational thinking, and self-reflection). Based on these findings, Greene (2013) argued that, compared with deontological morality, utilitarian morality is cognitively more costly and reflective, less emotional, and less intuitive.

However, more recent research undermines this interpretation; neural lesions (Christensen & Gomila, 2012; Koenigs et al., 2007), psychopathy, alexithymia (Patil & Silani, 2014), and some acute states of intoxication (Duke & Bègue, 2015; Perkins et al., 2013) can increase the tendency towards utilitarian moral judgments. Moreover, quick, intuition- and feel-based cognitive processing that does not require active step-by-step reasoning in working memory is a fundamental aspect of the human cognitive architecture. For instance, chess masters often rely on an intuitive "feel" for different moves and assessment of the "board as a whole" when there are too many move options to work through in working memory (e.g., Chassy & Gobet, 2011; Gobet & Chassy, 2009). Expert chess decision-making is largely a *feel-based* cognitive process, wherein the board configuration is compared to a vast knowledge-base of middle game positions encountered over thousands of hours of playing chess, solving chess problems, and reading chess literature (Chassy & Gobet, 2011; Gobet & Chassy, 2009). The same holds for many other domains, such as music, medical diagnostics, and poker, to name a few (Kahneman & Klein, 2009).

Thus, fast emotional reactions, intuitions, and feelings are equally "cognitive" as the processes involved in slow deliberative reasoning. In fact, automatic, feel-based cognition could be processing a larger number of bits than "slower" conscious cognition. Generally, the majority of human functioning is based on massive numbers of cognitive processes taking place outside of awareness, while conscious step-by-step calculations in working memory, mental "speech", and other similar phenomena are merely the tip of the iceberg.

The specific emotions of disgust, anger, and contempt have been linked to moral judgements (Steiger & Reyna, 2017). Anger motivates punishment (Gummerum et al., 2016): if somebody insults us, steals from us, or otherwise treats us badly, our anger motivates us to assert ourselves and defend our space, and to seek formal or informal punishment for the culprit. This signals to the transgressing person that their actions were costly. Disgust, too, has a significant, but contested, role in moral behaviors and judgments (Laakasuo et al., 2017; Tybur et al., 2013). Disgust may function as the gatekeeper of moral condemnation: things we find disgusting seem wrong and are therefore condemned. However, disgust is a complicated emotion. It is associated with the presence of pathogens (bacteria), but also with sexuality (as a reaction to unsuitable mates or non-normative forms of sexuality), and abstract issues (e.g., as a reaction to burning of the flag of one's home country).

A detailed analysis of different forms of disgust is beyond the scope of this chapter (see Tybur et al., 2013), but because of its relevance to our topic, we will briefly focus on the connection between disgust and morality. In our own studies, we have measured individual differences in trait *disgust sensitivity* (DS; Tybur et al., 2009) in relation to moral judgment. DS differs from *incidental* disgust experienced during a specific situation (i.e., the state of feeling disgusted). Feeling disgusted *per se* might not be associated with moral judgment (Landy & Goodwin, 2015), but DS is. In other words, the more people are disgust-sensitive,

the more likely they are to condemn different things. What is especially puzzling is the connection between sexual DS and completely non-sexual areas of moral judgment. The ostensible function of sexual disgust – in the evolutionary framework – is to guide mate selection and weed out potentially costly mating situations. Nevertheless, people more sensitive to sexual disgust are more conservative, careful about conventional norm violations, averse to drug use (Tybur et al., 2010), and less utilitarian (Laakasuo et al., 2017).

4. HUMAN COGNITIVE CATEGORIES SHAPED BY EVOLUTION

How can our understanding of the world be modelled? Although different sciences involved in studying cognition and knowledge-structures often disagree on the levels of explanation and analysis (Mitchell 2003; Horst 2016), almost all agree that our knowledge is nested into categories and concepts. These different knowledge forms can be mapped to procedural ("action"), semantic ("meaning"), and episodic ("event") memory systems (Tulving, 1985; Fletcher et al., 1999; Barrett, 2015: Farmer & Matlin, 2019).

Virtually all animals have procedural memory capabilities used for achieving various tasks (Tulving, 2002). However, humans in significantly higher extent create abstractions from their experiences and form semantic concepts (i.e., mental representations that have *meaning*). Semantic memory is crucial when making generalizations without actual experience (Binder & Desai, 2011). Our ability to categorize is linked to semantic memory, allowing us the use of conceptual knowledge beyond direct interaction with objects. This, in turn, makes human culture, science, religion, and art possible (Binder & Desai, 2011). That is, these cognitive properties enable us to recognize objects, create, and manipulate symbols to communicate with, and understand others, remember the past, and imagine the future. Declarative knowledge is necessary for essentially all uniquely human phenomena.

Some knowledge domains also seem to be more basic than others. The notion of "domain specificity" – cognitive structures operating on narrow and specific problems – describes this aspect of our cognition. Spelke (2000) has identified four basic "core knowledge systems" representing: 1) visuospatial structure, 2) objects and their interactions, 3) actions and goal directedness, and 4) numbers and relationships of ordering. These are processed via computational resources stored in sensorimotor programs and are crucial for the development of semantic domains (Spelke & Kinzler, 2007). Semantic domains are needed for different levels of intersubjectivity (for example representing the desires or emotions of others) and communication. This is the form of cognition that is commonly referred to as having a "theory of mind" (see section 2.).

EP and cognitive anthropology also study *natural categories* and intuitive biology, that is, innate, automatic, and domain-specific abilities to classify environmental stimuli into semantic categories (Atran, 2012; Boyer & Barrett, 2015). Small children can intuitively classify animals, plants, and rocks into their own categories. In learning their first words, infants already expect them to denote whole objects, rather than their parts. Similarly, when two objects collide, infants do not expect them to merge into one object (Boyer, 2018; Moll & Tomasello, 2010). Small children are, in general, very sensitive to a wide range of category violations. When children play, zebras do not eat lions, and trees do not walk and eat zebras (Boyer & Barrett, 2015). Moreover, in children's play, "a dog is still a dog" even when equipped with unusually big ears, a glued-on trunk, or other elements typically not associated with dogs (Gelman & Wellman, 1991).

Very early on, children have at least some level of understanding of different categories, the essential nature of different objects in these categories (e.g., what makes a certain individual organism an "animal"), and the causal relations between them. They expect

the "essence" of animals, instead of their external appearance, to be the reason why they behave in certain ways (Carey, 2009; Gallistel & Gelman, 2000; Hirschfeld & Gelman, 1994; Spelke & Kinzler, 2007). Children also expect animals to move by themselves, guided by the animal's own intentions and beliefs (Boyer, 2018) – that is, children perceive animals to have minds (see section 2).

This intuitive or innate understanding of the world is useful only if it reflects the structure of the outside reality to a good enough degree (Hoffman, 2019). Such understanding can be inaccurate, but not so inaccurate as to seriously impede survival. This knowledge has helped us during our evolutionary history to avoid making miscalculations and faulty predictions in a hostile and unpredictable environment, or when facing a complex new situation. But what about robots and other AIs? We did not confront them in the evolutionary savannah or Pleistocene forest. Moreover, they are a relatively new phenomenon even to modern humans. This means that robots and AIs do not belong in any evolution-given natural category, or necessarily even in any cultural one. Still, similar to other animate agents, we may perceive them as having internal states; intentions and even beliefs. They can stimulate and bias our cognition in unpredictable ways. The way in which we classify them may significantly affect our ethical and moral cognitions concerning them.

4.1 Illustration: Tools as the First Cognitive Category Shaped by Technology

Tools are a universal human cognitive category with deep evolutionary roots. Tools can be defined as commonly hand-held objects that make it easier to carry out specific tasks. Modern chimpanzees and even ancient Australopithecines used tools similar to early hominid tools, such as chipped stones (Stanford et al., 2011). Initially, all three species used unmodified tools (objects found in nature); but over time the tools used by *Homo sapiens* became self-made and more sophisticated. The differences between modern humans and

other modern primates in their tool use reflect evolutionary differences in multiple traits between the species, such as hand-eye coordination, causal reasoning, social intelligence, learning, and language (Vaesen, 2012).

In human evolution, there are two long and specific periods where two types of tool-creating cultures existed; the Oldowan periods (started 2.5 million years ago) and the Acheulean periods (1.75 mya), both of which lasted for about a million years. These periods are commonly described as periods of archeological boredom, since during them, large quantities of very similar stone tools were produced. These time periods, however, seem to have been long enough to act as selection pressures for human cognition. Brain imaging studies have revealed that an area on the left side of the brain - left anterior supramarginal gyrus, aSMG - responds in a species-specific and unique way to images of tools; just mimicking tool use or hearing the tool's sound is enough to activate aSMG (Orban & Caruana, 2014). In addition, Uomini and Meyer (2013) observed similar brain activation when subjects were silently thinking of words starting with a given letter and when they were knapping an Acheulean flint axe, suggesting that tool-making and language share a basis in more general human capacities for complex, goal-directed action.

Furthermore, different types of brain lesions are associated with consequences on tool use. After certain types of brain damage, otherwise normally behaving patients may no longer understand how to bend their arms to use a hammer: some do not understand what to use a hammer for, and some may have lost the whole concept of "hammer" from their minds (Baumard et al., 2014). This suggests that we have specialized systems in our brains for tools,

and if they are malfunctioning, we lose our special ability to use, or even think flexibly about, tools. Studies also show that using tools has a cascading effect on our species' survival. Intelligence allowed us to use tools, tools helped us gather and process more food, food helped nourish our brains, better nourished brains allowed more intelligence which again allowed more and improved tools, taking us from stone age to today (Flinn et al, 2006; Ko, 2016). Thus, the cognitive category of tools seems to have deep biological origins in our evolutionary history.

Machines are certain kinds of tools recently introduced in human cultural history; i.e., they are a cultural category distinct from purely cognitive categories. The category of "machine" is now rooted in our collective thinking, and most of us have a basic understanding of what various different machines are used for – even if we could not give a detailed breakdown of their functionality. We generally view machines as devices, or tools, to achieve some goal: dishwashers are used for washing dishes, cars for transportation, and microwaves for heating food. However, machines such as robots (even without intelligence) and intelligent programs, fool us into thinking they have minds, thus challenging our cognition in ways previously unseen in our evolutionary history. A hypothesis can be made that robots, being able to move autonomously, and intelligent programs being able to reason, make decisions, sometimes talk, and unlike any other inanimate objects, activate some automatic (instinctive, sub-conscious, evolutionary programmed) cognitive reactions classifying those objects as human-like, even though we know they are just artifacts.

5. BIO-CULTURAL HUMANS AND THE NEW ONTOLOGICAL CATEGORY

Kahn et al. (2011) suggest that artificial agents form a *new ontological category*: robots and AIs are something entirely novel in the natural and cultural history of our planet. Thus, when interacting with robots we must rely on intuitions that evolved in the absence of robots. This may result in various forms of categorical confusions and misinterpretations when dealing with robots or other AIs.

The *uncanny valley effect* (UVE) is a classic example of perceptual categorical confusion (Mori, 1970) – but not specific towards robots. The UVE occurs when the appearance of a robot (or any non-human agent) passes a certain threshold of similarity to humans, and we become repulsed by it. The actual mechanisms underlying the UVE are still unclear (Palomäki et al., 2018). Nevertheless, by watching YouTube videos made of the humanoid robot Sophie, one can perhaps affirm that there is something creepy about her (see, e.g., CNBC, 2016). Does Sophie belong in the category of *animate living objects* (like other people), or *inanimate objects* (like dolls)? She might even be categorized as a *tool* if she were viewed as a robot with specific functions to perform.

In the *biocultural* view of humans, biology and culture are perceived as intertwined entities feeding into one another (Fuentes, 1999; Donald, 2002; Richerson & Boyd, 2005). Various religions across the world can be seen as products of this dynamic interaction (Atran & Norenzayan 2004; Geertz, 2010; Sørensen 2004), and within religions there are many examples of new cultural categories having emerged that utilize our evolutionarily old mechanisms. For example, it is well documented that animistic cultures and tribal communities had (and have) rich spiritual conceptual worlds, with beliefs and ritual practices that held natural objects as animate and essential agents (Atran, 2002; Boyer, 2001; Lawson & McCauley, 1990; McCauley & Whitehouse, 2005). In the animistic worldview, everything is connected; minds and mental states are attributed to natural objects, different kinds of

spirits, deities, gods, and ancestors similar to humans: they can set goals, have intentions, and feel emotions. Early animistic cultures had moralizing gods, varying in the degree to which they cared about the morality of their followers (Boyer, 2001; Purzycki et al., 2016; Willard & McNamara, 2016). These beliefs about supernatural agents spread through ancestral rituals and migration (Norenzayan et al., 2015). From the viewpoint of cognitive science, animism is an example of people associating human-like agency either to things that do not act in any way (such as rocks and trees), or that do not necessarily have any clear agency (e.g., natural events such as rain seen as "caused" by spirits).

Technological animism is a new cultural concept of personhood that is emerging from the interaction between fiction, robotics, and different cultural models of agency (Richardson, 2018). Technological animism has already had an impact in human-robot interaction. For example, Japanese roboticists radically differ from their Euro-American collegues in their use of animistic elements from Japanese Buddhism and Shintoism to support a cultural narrative of robots as friends instead of enemies (Coeckelbergh, 2013; Jensen & Blok, 2013). However, even in Western countries, children tend to associate human-like emotional and cognitive capabilities to robots, prompting researchers to instruct parents to explicitly teach their children to call a robot "it" (Shellenbarger, 2019). Without explicit cultural training or education, if they manage to avoid the uncanny valley, robots appear to inspire seemingly animistic thinking.

To be clear, our claim is that artificial agents may activate similar cognitive processes related to agency, mind perception, and morality as seen in "animistic" interpretations of non-human objects or natural phenomena.We do not mean to claim that robots and AIs will induce religious or spiritual behavior in humans (however, see Harris, 2017, for a report on the first "church of AI"). Mind perception may even be easier in the case of robots and AIs than, for example, other animals, rocks, or natural events, as both robots and AIs can be made

intentionally more human-like. In addition, we can concretely observe robots (i.e., AIs with *bodies*) in action, and see that their behavior causes specific things to happen in the world. Thus, from the cognitive perspective, robots are by nature closer to humans or other animals than to inanimate objects such as rocks or trees. However, the logical, probabilistic computations whereby AIs function are often opaque and even counterintuitive to humans (see Rode et al., 1999 on human difficulties with probabilities). We will return to this mismatch between the way humans think of agency and the kinds of agency robots and AIs actually have in Part 2, where we will utilize the theories presented here in detail.

6. CONCLUSION

In this first of two chapters, we shortly summarized the basics of evolutionary and moral psychology. We also reviewed the basics of how human categorization of natural environments may run into issues with novel technologies. We discussed how emotions and reason have a complex interplay and give rise to our moral cognitive judgments, and how previous technological stages in human evolution have influenced the development of our cognitive system and given us the concept of tools. We concluded the chapter by introducing the concept of the *new ontological category*, and discussed how we do not have the evolutionary capabilities to deal with robots and other intelligent information processing systems intuitively. Given that it took us two million years or so, to evolve the concept of tools, it seems that we only have cultural solutions to the new moral problems facing us in the technological domain. One of the main cognitive mechanisms that we currently utilize in our interaction with robots and AIs is the mind perception mechanism, which is nonetheless constantly fooled into projecting minds to where there are none, at least for now.

REFERENCES

Allhoff, F., Lin, P., & Steinberg, J. (2011). Ethics of human enhancement: an executive summary. *Science and engineering ethics*, 17(2), 201-212.

Alicke, M. D. (2012). Self-Injuries, Harmless Wrongdoing, and Morality. *Psychological Inquiry, 23,* 125-128. https://doi.org/10.1080/104780X.2012.666720

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. Oxford University Press.

Atran, S. (2012). Psychological orgins and cultural evolution of religion. In R. Sun (Ed.), *Grounding Social Sciences in Cognitive Sciences* (pp. 209–238). MIT Press.

Atran, S., & Norenzayan, A. (2004). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences*, *27*(6), 713–730.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199348305.001.0001

Baumard, J., Osiurak, F., Lesourd, M., & Le Gall, D. (2014). Tool use disorders after left brain damage. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00473

Baumeister, R. F., & Leary, M. R. (1995). The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological Bulletin*, *117*(3), 497–529.

Bentham, J. (1816). *Chrestomathia*. Edinburgh: William Tait

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. Basic Books.

Boyer, P. (2018). *Minds make societies: How cognition explains the world humans create*. Yale University Press.

Boyer, P., & Barrett, H. C. (2015). Intuitive Ontologies and Domain Specificity. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 161–174). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119125563.evpsych105

Breggin, P. R. (2015). The biological evolution of guilt, shame and anxiety: A new theory of negative legacy emotions. *Medical Hypotheses, 85(1), 17–24.* https://doi.org/10.1016/j.mehy.2015.03.015

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Castelo, N., Schmitt, B., & Sarvary, M. (2019). Human or Robot? Consumer Responses to Radical Cognitive Enhancement Products. *Journal of the Association for Consumer Research*, *4*(3), 217–230. https://doi.org/10.1086/703462

Chassy, P., & Gobet, F. (2011). A Hypothesis about the Biological Basis of Expert Intuition. *Review of General Psychology*, *15*(3), 198–212. https://doi.org/10.1037/a0023958

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Coeckelbergh, M. (2013). *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*. Springer Science & Business Media.

Cohen-Chen, S., Halperin, E., Saguy, T., & van Zomeren, M. (2014). Beliefs About the Malleability of Immoral Groups Facilitate Collective Action. *Social Psychological and Personality Science*. 5(2), 203–201. https://doi.org/10.1177/1948550613491292

Curry, O. S., Chesters, M. J., Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of "morality-as-cooperation" with a new questionnaire. *Journal of Research in Personality, 78,* 106-124. https://doi.org/10.1016/j.jrp.2018.10.008

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience* 7(3). https://doi.org/10.1080/17470919.2011.614000

Donald, A. (2002). *A mind so rare: The evolution of human consciousness*. W. W. Norton and Company.

Doris, John, Stich, Stephen, Phillips, Jonathan and Walmsley, Lachlan. (2020). in Edward N. Zalta(ed.) Moral Psychology: Empirical Approaches, *The Stanford Encyclopedia of Philosophy.*

Duke, A. A., & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition*, *134*, 121–127. https://doi.org/10.1016/j.cognition.2014.09.006

Dunbar, R. (2014). *Human evolution*. Pelican Books.

Feinberg, M., Willer, R., & Keltner, D. (2011). Flustered and Faithful: Embarrassment as a Signal of Prosociality. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/a0025403

Fjell, C. D., Cherkasov, A., Hilpert, K., Jenssen, H., Waldbrook, M., Mullaly, S. C., Volkmer, R., & Hancock, R. E. W. (2008). Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chemical Biology*. 4(1). 65-74. https://doi.org/10.1021/cb800240j

Fletcher, P., Büchel, C., Josephs, O., Friston, K., & Dolan, R. (1999). Learning-related Neuronal Responses in Prefrontal Cortex Studied with Functional Neuroimaging. *Cerebral Cortex*, *9*(2), 168–178. https://doi.org/10.1093/cercor/9.2.168

Flinn, M. V., Geary, D. C., & Ward, C. V. (2005). Ecological dominance, social competition, and coalitionary arms races: Why humans evolved extraordinary intelligence. *Evolution and Human Behavior*, *26*(1), 10–46. https://doi.org/10.1016/j.evolhumbehav.2004.08.005

Francis, R. C. (1990). Causes, proximate and ultimate. *Biology & Philosophy*, *5*(4), 401–415. https://doi.org/10.1007/BF02207379

Fuentes, A. (1999). *Evolution of human behavior*. Oxford University Press.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65. https://doi.org/10.1016/S1364-6613(99)01424-2

Geertz, A. W. (2010). Brain, Body and Culture: A Biocultural Theory of Religion. *Method & Theory in the Study of Religion*, *22*(4), 304–321. https://doi.org/10.1163/157006810X531094

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213–244. https://doi.org/10.1016/0010-0277(91)90007-Q

Gobet, F., & Chassy, P. (2009). Expertise and Intuition: A Tale of Three Theories. *Minds and Machines*, *19*(2), 151–180. https://doi.org/10.1007/s11023-008-9131-5

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Greene, J. (2007). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3*. MIT Press.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Guillette, S. (2019, December 6). Your new lifeguard may be a robot. *Verizon*.

Grüter, M., von Kriegstein, K., Dogan, Ö., Giraud, A., Kell, C. A., Grüter, T., Kleinschmidt, A,. & Kiebel, S. J. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences*. 105(18), 6747-6752. https://doi.org/10.1073/pnas.0710826105

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the

perpetrator and compensation of the victim. *Journal of Experimental Social Psychology*, *65*, 94–104. https://doi.org/10.1016/j.jesp.2016.04.004

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 21.

Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology, 31(1).* https://doi.org/10.1111/j.1559-1816.2001.tb02489.x

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002.

Haidt, J. (2008). Morality. *Perspectives on Psychological Science*, *3*(1), 65–72.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and Below Left–Right: Ideological Narratives and Moral Foundations. *Psychological Inquiry*, *20*(2–3), 110–119. https://doi.org/10.1080/10478400903028573

Harris, M. (2017, November 15). Inside the First Church of Artificial Intelligence. *Wired*.

Helkama, K. (2009). *Moraalipsykologia: Hyvän ja pahan tällä puolen*. Edita Publishing Oy.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). *Mapping the mind*. Cambridge University Press.

Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. W. W. Norton and Company.

Horst, S. (2016). *Cognitive pluralism*. The MIT Press. https://muse.jhu.edu/book/46963

Jensen, C. B., & Blok, A. (2013). Techno-animism in Japan: Shinto Cosmograms, Actor-network Theory, and the Enabling Powers of Non-human Agencies. *Theory, Culture & Society*, *30*(2), 84–115. https://doi.org/10.1177/0263276412456564

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. https://doi.org/10.1016/j.cognition.2014.10.005

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. *Proceedings of*

the 6th International Conference on Human-Robot Interaction - HRI '11, 159. https://doi.org/10.1145/1957656.1957710

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Keesing, R., & Strathern, A. J. (1997). *Cultural Anthropology: A Contemporary Perspective (3rd Edition)*. Wadsworth Publishing.

Keltner, D., Haidt, J., & Shiota, M. N. (2006). Social functionalism and the evolution of emotions. In M. Schaller, J. A. Simpson, & D. T. Kendrick (Eds.), *Evolution and social psychology* (pp. 115–142). Psychology Press.

Ko, K. H. (2016). Origins of human intelligence: The chain of tool-making and brain evolution. *Anthropological Notebooks*, 5–22.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911. https://doi.org/10.1038/nature05631

Koverola, M., Kunnari, A., Palomäki, J., Drosinou Marianna., & Laakasuo Michael. (in press). Moral Psychology of Sex Robots: an experimental study – How Pathogen Disgust is associated with interhuman sex but not interandroid sex. *PALADYN – Journal of Behavioral Robotics*.

Laakasuo, M., Drosinou, M., Koverola, M., Kunnari, A., Halonen, J., Lehtonen, N., & Palomäki, J. (2018). What makes people approve or condemn mind upload technology? Untangling the effects of sexual disgust, purity and science fiction familiarity. *Palgrave Communications*, *4*(1), 1–14. https://doi.org/10.1057/s41599-018-0124-6

Laakasuo, M., Köbis, N., Palomäki, J., & Jokela, M. (2018). Money for microbes-Pathogen avoidance and out-group helping behaviour. *International Journal of Psychology*, *53*, 1–10. https://doi.org/10.1002/ijop.12416

Landy, J. F., & Goodwin, G. P. (2015). *Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence*. *10*(4), 518–536.

Lawson, E. T., & McCauley, R. N. (1990). *Rethinking religion: Connecting cognition and culture*. Cambridge University Press.

Levy, D. (2007). *Love and sex with robots*. Harper Collins.

Lee, S.-H., & Wolpoff, M. H. (2003). The pattern of evolution in Pleistocene human brain size. *Paleobiology*, *29*(2), 186–196.

Lewis-Williams, D. (2002). *The mind in the cave: Consciousness and the origins of art*. Thames & Hudson.

Li, L., Lv, Y., & Wang, F. Y. (2016). Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, *3*(3), 247-254.

Licata, M., Zietlow, A.-L., Träuble, B., Sodian, B., & Reck, C. (2016). Maternal Emotional Availability and Its Association with Maternal Psychopathology, Attachment Style Insecurity and Theory of Mind. *Psychopathology*, *49*(5), 334–340. https://doi.org/10.1159/000447781

Loughnan, S., Haslam, N., Murnane, T., Vaes, J., Reynolds, C., & Suitner, C. (2010). Objectification leads to depersonalization: The denial of mind and moral concern to objectified others. *European Journal of Social Psychology*, *40*(5), 709–717. https://doi.org/10.1002/ejsp.755

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). *AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma.*

Mangasarian, O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern Recognition Via Linear Programming: Theory And Application To Medical Diagnosis.

McAuliffe, W. H. B. (2019). Do emotions play an essential role in moral judgments? *Thinking & Reasoning*, *25*(2), 207–230. https://doi.org/10.1080/13546783.2018.1499552

McCauley, R. N., & Whitehouse, H. (2005). New frontiers in the cognitive science of religion. *Journal of Cognition and Culture*, *5*, 1–13.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Mill, J. S. (1861). Utilitarianism. *Fraser's Magazine, 64, 391–406; 525–534; 659–673*.

Mitchell, S. D. (2003). *Biological Complexity and Integrative Pluralism* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511802683

Moll, H., & Tomasello, M. (2010). Infant cognition. *Current Biology*, *20*(20), R872–R875.

Monroe, A., Guglielmo, S., & Malle, B. (2012). Morality Goes Beyond Mind Perception. *Psychological Inquiry, 23,* 179-184. https://doi.org/10.1080/1047840X.2012.668271

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Norenzayan, A., Shariff, A. F., Gervais, W. M., Willard, A. K., McNamara, R. A., Slingerland, E., & Henrich, J. (2016). The cultural evolution of prosocial religions. *Behavioral and Brain Sciences*, *39*, e1. https://doi.org/10.1017/S0140525X14001356

Orban, G. A., & Caruana, F. (2014). The neural basis of human tool use. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00310

Palomäki, J., Kunnari, A., Drosinou, M., Koverola, M., Lehtonen, N., Halonen, J., Repo, M., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon*, *4*(11). https://doi.org/10.1016/j.heliyon.2018.e00939

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00501

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., Williams, S. C. R., & Ettinger, U. (2013). A dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General*, *142*(3), 612–620. https://doi.org/10.1037/a0030256

Pinker, S. (1994). *The language instinct*. William Morrow and Company.

Pinker, S. (1997). *How the mind works*. W. W. Norton and Company.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. Penguin Books.

Poon, S. H., Jondeau, E., & Rockinger, M. (2007). *Financial modelling under non-Gaussian distributions*. Springer Science & Business Media.

Purzycki, B. G., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., Xygalatas, D., Norenzayan, A., & Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature*, *530*(7590), 327–330. https://doi.org/10.1038/nature16980

Richardson, K. (2016). Technological Animis: The Uncanny Personhood of Humanoid Machines. *Social Analysis*, *60*(1), 110–128. https://doi.org/10.3167/sa.2016.600108

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rigby, J., Conroy, S., Miele-Norton, M., Pawlby, S., & Happé, F. (n.d.). Theory of mind as a predictor of maternal sensitivity in women with severe mental illness. *Psychological Medicine*, *46*(9), 1853–1863.

Rode, C., Cosmides, L., Hell, W., & Tooby, J. (1999). When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition*, *72*(3), 269–304. https://doi.org/10.1016/S0010-0277(99)00041-4

Saxe, R., & Baron-Cohen, S. (2006). Editorial: The neuroscience of theory of mind. *Social Neuroscience*, *1*(3–4), 1–9. https://doi.org/10.1080/17470910601117463

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Shellenbarger, S. (2019, August 26). Why We Should Teach Kids to Call the Robot 'It.' *The Wall Street Journal*. https://www.wsj.com/articles/why-kids-should-call-the-robot-it-11566811801

Shultz, S., Nelson, E., & Dunbar, R. I. M. (2012). Hominin cognitive evolution: Identifying patterns and processes in the fossil and archaeological record. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2130–2140. https://doi.org/10.1098/rstb.2012.0115

Sidgwick, H. (1874). *The Methods of Ethics.* London: MacMillan & CO

Singh, M., & Naveen, B. P. (2014). Molecular Nanotechnology: A new avenue for Environment Treatment. *Journal of Environmental Science, Toxicology And Food Technology*, *8*(1), 93-99.

Sinnott-Armstrong, W. (ed. 2014). *Moral psychology, Vol 4: Free will and moral responsibility*. MIT Press.

Sørensen, J. (2004). Religion, evolution, and an immunology of cultural systems. *Evolution and Cognition*, *10*(1), 61-73.

Spelke, E. S. (2000). Core knowledge. *The American Psychologist*, *55*(11), 1233–1243. https://doi.org/10.1037//0003-066x.55.11.1233

Spelke, Elizabeth S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Springer, P. J. (2013). *Military robots and drones: A reference handbook*. ABC-CLIO.

Stanford, C., Allen, J. S., & Antón, S. C. (2011). *Biological anthropology*. Pearson Education.

Steiger, R. L., & Reyna, C. (2017). Trait contempt, anger, disgust, and moral foundation values. *Personality and Individual Differences*, *113*, 125–135. https://doi.org/10.1016/j.paid.2017.02.071

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00229

Tooby, J., & Cosmides, L. (1992). The Psychological Foundations of Culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford University Press.

Tooby, J., & Cosmides, L. (2005). Conceptual Foundations of Evolutionary Psychology. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 5–67). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470939376.ch1

Tsoukalas, I. (2018). Theory of Mind: Towards an Evolutionary Theory. *Evolutionary Psychological Science*, *4*(1), 38–66. https://doi.org/10.1007/s40806-017-0112-x

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tybur, J. M., Lieberman, D., & Griskevicius, V. (2009). Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology*, *97*(1), 103–122. https://doi.org/10.1037/a0015474

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65–84. https://doi.org/10.1037/a0030778

Tybur, J. M., Merriman, L. A., Hooper, A. E. C., McDonald, M. M., & Navarrete, C. D. (2010). Extending the Behavioral Immune System to Political Psychology: Are Political Conservatism and Disgust Sensitivity Really Related? *Evolutionary Psychology*, *8*(4), 147470491000800420. https://doi.org/10.1177/147470491000800406

Uomini, N. T., & Meyer, G. F. (2013). Shared Brain Lateralization Patterns in Language and Acheulean Stone Tool Production: A Functional Transcranial Doppler Ultrasound Study. *PLoS ONE*, *8*(8). https://doi.org/10.1371/journal.pone.0072693

Vaesen, K. (2012). The cognitive bases of human tool use. *The Behavioral and Brain Sciences*, *35*(4), 203–218. https://doi.org/10.1017/S0140525X11001452

Voyer, B. G., & Tarantola, T. (2017). Toward a Multidisciplinary Moral Psychology. In B. G. Voyer & T. Tarantola (Eds.), *Moral Psychology: A Multidisciplinary Guide* (pp. 1–3). Springer International Publishing. https://doi.org/10.1007/978-3-319-61849-4_1

Willard, A. K., & McNamara, R. A. (2019). The Minds of God(s) and Humans: Differences in Mind Perception in Fiji and North America. *Cognitive Science*, *43*(1), e12703. https://doi.org/10.1111/cogs.12703

Zeytinoglu, S., Calkins, S. D., & Leerkes, E. M. (2018). Maternal emotional support but not cognitive support during problem-solving predicts increases in cognitive flexibility in early childhood. *International Journal of Behavioral Development*. https://doi.org/10.1177/0165025418757706