Moral Psychology and Artificial Agents (Part Two) : The Transhuman Connection

Laakasuo, Michael

IGI Global
2021

Laakasuo , M , Sundvall , J R I , Berg , N A , Drosinou , M-A , Herzon , V , Kunnari , A , Koverola , M , Repo , M , Saikkonen , T & Palomäki , J 2021 , Moral Psychology and Artificial Agents (Part Two) : The Transhuman Connection . in S J Thompson (ed.) , Machine Law, Ethics and Morality in the Age of Artificial Intelligence . Advances in Human and Social Aspects of Technology (AHSAT) , IGI Global , Hershey, PA , pp. 189-204 . https://doi.org/10.4018/978-1-7998-4894-

http://hdl.handle.net/10138/337163
https://doi.org/10.4018/978-1-7998-4894-3.ch011

acceptedVersion

Moral Psychology and Artificial Agents (Part 2):

The Transhuman Connection

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 2): The Transhuman Connection. *Machine Law, Ethics and Morality in the Age of Artificial Intelligence*. Steven Thompson (ed). New York: Igi Global.

INTRODUCTION

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 towards the world outside our minds. In Part 1, we showed how evolution equipped us with such capacities as mind perception (understanding other minds), tool use, and the emotion of disgust (pathogen avoidance)[1]. We concluded the previous chapter by analyzing the concept of the New Ontological Category. The New Ontological Category, i.e., Robots, AIs, and other forms of intelligent technologies which are neither alive nor inanimate (but non-alive and animate), did not exist during our evolution. We therefore do not have an intuitive understanding of them the same way we have of humans, animals, plants, rocks, and inanimate matter. This fundamental observation, stemming from the basic findings in evolutionary psychology, then makes it salient that there are bound to be odd and unexpected clashes between new "animate" technologies and the human moral cognitive system. Human understanding of the world divides the world into epistemic categories, of inanimate nonliving and animate living, but it has no conceptual category naturally corresponding to the *new ontological category* which dilutes these categories to "animate non-living."

Here we show how the fact that we do not have intuitive understanding of the *new ontological category* unwinds in unexpected and unpredictable ways in the recent moral psychological literature focused on understanding human moral psychology in new contexts.

1 We remind the reader that moral psychology is a descriptive science, not normative philosophy. Naturally, there are more emotions than just disgust associated with human moral behavior. However, disgust is probably the most extensively studied emotion in the field of moral psychology and is therefore one of the main things we cover here.

We will show how category violations that happen between humans and machines, and between minded and non-minded entities, results in inconsistent moral judgments, which are not explained by existing moral psychological theories. We start with the definition of transhumanism – a long standing philosophical project that aims to redefine what humanity is – then discuss the categorical clashes between our cognition, robots, and human enhancement technologies (Thompson, 2014). The reason for covering transhumanism here is that transhumanism is a philosophy that fundamentally blurs the categories of humans and machines. Transhumanistic technologies, such as cognitive enhancement, can create similar confusions to our moral psychological apparatus, as do robots, AIs, and other information processing technologies. In a sense, transhumanism pulls humans into the New Ontological Category as well, since it fundamentally sees humans as information processing systems that can and should be integrated with AIs and machines. This is guaranteed to create moral cognitive clashes, as we will show.

After covering the basics of transhumanist philosophy (Section 2), the moral psychology of robotics (Section 3.1), and the moral psychology of transhumanist technologies (Section 3.2), we then move on to discuss the limits of present-day moral psychological theories (Section 4). These are limits made obvious by the *new ontological category*; we also suggest some directions for future studies (Section 5). Finally, we conclude by summarizing the lessons learned (Section 6).

2. DEFINING TRANSHUMANISM

The term “transhumanism” was originally defined by biologist and first UNESCO director, Julian Huxley, as a belief in the possibility of “man remaining man, but transcending himself,

by realizing the new possibilities of and for his human nature" (1957). Whereas Huxley emphasized both the spiritual and communal aspects of this enterprise, the term was later adopted by thinkers focused on the technological aspects of human improvement; that is, human cognitive and physical enhancement or alteration going beyond our normal limits. Today, *transhumanism* is an umbrella term for philosophical, religious, aesthetical, social and political movements, engineering, various research projects, worldviews, and lifestyles claiming that: a) the current state of humanity is not the endpoint of its evolution, and b) humanity can (and maybe should) take conscious action to guide its own evolution through technological means (O'Connell, 2018; Thompson, 2014; See also Lin et al., 2014a; 2014b).

While there is no strict universal transhumanist moral code, transhumanist arguments do have strong utilitarian leanings (Bostrom, 2005; Frohlich, 2015; More, 2010; Sotala & Gloor, 2017). Critiques of transhumanism have centered on broad concerns about loss of the meaning of life, sanctity of the body, or essential humanity (Kass, 2003); or claimed that blind faith in technological progress is equivalent to believing in a benevolent God who will *ex machina* solve our problems (Burdett, 2014). Some fear that transhumanistic technologies and visions might alter our social interactions by making them more superficial and instrumental (Frischmann & Selinger, 2018). Further criticism has focused on the potential damaging socioeconomic effects of transhumanism, suggesting it might widen the gap between the haves and the have-nots, coerce eugenics, or lead to a permanent division of humanity into a master and a slave class (Moravec, 1988). To all of these concerns, the transhumanist response might be summed up in Russell Blackford's words:

> The concern is essentially a matter of social justice, a problem that modern societies must, and do, wrestle with continually, within real-world economic and political constraints. Responses to the problem might vary from redistribution of the wealth that enables differential access in the first place, prohibition or

> limitation of enhancement technologies in the interests of fairness, or steps to make at least some genetic technologies – those relating to health and longevity – as widely available as possible. (2003)

The key areas of transhumanist interest[2] are different forms of cognitive and physical enhancement (e.g., drugs, implants, and other technological aids to performance), life extension technologies (e.g., cryonics, cloning, "mind-uploading," eugenics, and gene therapies against aging), and new sentient "life forms" (e.g., radically augmented humans, robotics, brain-computer symbiotes, "whole brain emulations"). Attempts to increase human performance and longevity, or to change our very nature, also relate to how we categorize things. How much can a human being be altered before they are no longer (categorized as) human, and what does this mean in terms of their moral status? Technologies that make people radically better at something through invasive means also raise questions of harm and fairness, making above-baseline human enhancements novel challenges for our moral cognition (Thompson, 2014; Lin et al, 2014a, 2014b).

One central aspiration of transhumanism is to increase the levels of human general intelligence, for instance, through eugenics, gene editing, and other biotechnological innovations (Bostrom, 2014). However, transhumanists usually talk about eugenics from an ahistorical perspective, where they are concerned about improving the quality of human life. Their perspective on eugenics is then radically different from the perspective of the general population – a theme we will return to in section 4 (Future Studies).

MORAL PSYCHOLOGICAL PERSPECTIVES ON ROBOTICS AND TRANSHUMANISM

Here, we will present recent empirical evidence from moral psychology showing how human moral cognition and intelligent technologies collide with unexpected results. We will first

2 See Transhumanist FAQ 3.0. http://www.whatistranshumanism.org

review recent findings on moral psychology of robotics, and then continue to present some intriguing novel findings from research studying attitudes towards transhumanistic technologies.

3. 1 Human Universals in the Moral Psychology of Robots?

In Part 1, we argued that mind perception is a crucial part of morality: we need to perceive something as an agent to hold it morally responsible, and to perceive something as having an inner experiential world to see it as also having moral rights. Here, we focus on the ways mind perception and moral judgment interact, and review studies on human preferences regarding how artificial agents act in moral situations.

In a rare study of robots as the victims (rather than perpetrators) of harm, Ward et al. (2013) turned the morality of machines upside down, suggesting that the link between mind perception and morality works both ways. Machines and corpses by definition are non-living and have no true moral interaction with humans. Robots that were intentionally abused were attributed *more* cognitive capacities than robots that were untouched; while (conscious) humans who were intentionally harmed were attributed *less* cognitive capacities. Thus, in the case of harm-causing abuse, mind perception depends on the status of the abused agent.

In a similar vein, several age groups of children between 7 and 15 years old thought it was equally morally wrong to maltreat the robot dog AIBO and a real dog (Melson et al., 2009). These findings align with the Theory of Dyadic Morality *(*TDM) (see Part 1, section 3), particularly the idea of "dyadic completion." That is, the TDM's proposed "dyadic loop" has three elements that form the template of a moral violation: a perpetrator, a victim, and harm. "Dyadic completion" means that people may infer all three elements to be present (and, thusly, that moral violation has occurred) when, in reality, only one or two elements are

present (Schein & Gray, 2018). In other words, merely *perceiving* a typical morally questionable action (e.g., hitting something) can be enough to "inject mind" into the situation.

In another recent paper, Bigman and Gray (2018) argue that people have earlier denied full moral status to children[3], animals, and even to other races (see below section 4 on dehumanization), and the same might be true for machines. Machine *agency* and responsibility may be linked to the extent that people perceive the machines as minded entities. In six of their studies, the authors found that people are averse to machines making moral decisions: people prefer machines not to decide on matters of life and death. This aversion arguably stems from thinking that machines lack a mind. This aversion is also not easy to overcome, as it persists even if the machines are described as having expertise or capacity for mental experience.

However, people do sometimes have clear opinions on what a robot moral agent should do in matters of life and death, even if they are averse to the very idea. Awad et al. (2018) argue, based on an extensive attitude survey on autonomous vehicles (AVs), that the dream of universal machine ethics is not doomed, since there are points of relative agreement between broad geographical regions. The authors studied how people react to utilitarian decisions made by AVs with cross-cultural data (over 40 million answers to moral dilemmas achieved through an online gamified survey platform). They wanted to know what types of road users, from domestic animals to individuals with high social status, survey participants would be willing to sacrifice. People generally preferred the utilitarian option of saving the most people possible. Additionally, people preferred to save humans over animals, and young over old. Demographic factors did not have an impact, but three cultural clusters were detected: the occidental, oriental, and southern clusters. In the oriental cluster, people had a lower preference to save young over old people. In the southern cluster (mostly Latin

3 See United Nations Global Issues. https://www.un.org/en/sections/issues-depth/children/

America), people had a lower preference to save humans over animals, but were more willing to save individuals with high status over those with lower status. The findings are in line with Evolutionary Psychology (EP, See Part 1) theory: human moral judgments are relatively universal, with some regional variations around a common "core" of moral thought.

However, AVs also reveal a darker side of the seeming universal utilitarian preference. Bonnefon et al. (2016) scrutinized the AV utilitarian dilemmas, finding that people are willing to recommend utilitarian vehicles for others, but for themselves prefer AVs that would protect them at all costs. Assuming both utilitarian and self-protective cars were on the market simultaneously, few people expressed willingness to buy the car with the utilitarian algorithm. Additionally, a majority of Bonnefon et al.'s participants, when asked, opposed the idea of regulations that would force AVs to be utilitarian (2016). Such contradictions in morals are not new when it comes to humans. However, it is interesting how clearly the idea of automated moral agents highlights these contradictions. When it comes to an actual autonomous moral decision-maker a person can buy, one cannot demand one kind of "moral car" for themselves and another kind for others. Moreover, the moral choices of these AVs would be pre-specified rather than left to be decided by people (with little time) – Bonnefon et al.'s results suggest that people do not favor such prespecification, even if it saved more lives on aggregate (2016). The question of truly utilitarian cars is essentially the following: Would you put your life in the hands of something, that may decide it is better for you to die to prevent a larger loss of life, and with which you cannot negotiate?

Waytz et al. (2014) studied the effects of enhanced anthropomorphism on the perception of AVs. Car manufacturers design their products to represent something they presume potential buyers desire in terms of driving comfort, power, and aesthetics. Waytz et al. went further by adding a name, gender and voice to one of their AV simulators (2014). This anthropomorphic AV simulator was rated more trustworthy and likeable than a non-

autonomous vehicle simulator or an AV simulator without any anthropomorphic features. Naming cars is not exactly new: four-time Formula 1 world champion, Sebastian Vettel, has named his racing cars with seductive female names (such as Kinky Kylie or Hungry Heidi). Boats are also traditionally named after people. However, these are instances of people naming things they like and control. A manufacturer adding a name and a voice (similar to the Alexa voice assistant) to a widely distributed product is entirely different and potentially riskier. Waytz et al. (2014) show how easy it is to manipulate people into seeing agency in, or feeling trust towards, lifeless objects, through simple manipulations that do not bear on the AV's primary function. A name and a voice may make an AV more approachable to humans, but increased trust without an increased understanding of how the machine works is a risky combination.

Malle et al. (2019) focused on different types of agents and moral responses these agents evoke in a military context. Participants were requested to judge the actions of a human military pilot, an autonomous drone, or an aircraft with artificial intelligence. The agents were to either carry out an attack on terrorists while risking the life of an innocent child wandering in the target area, or to cancel the attack to protect the child, which, in turn, risked a terrorist strike. Participants treated all the agents as more or less morally responsible: even the autonomous drone was condemned by half of the respondents. When people were asked what the agent should do, launching the strike was generally considered the better option for each of the agents. People thus imposed similar norms on all three agents. However, people morally evaluated a human and artificial agent's decision in an identical dilemma differently, blaming the human pilot who cancelled the attack significantly more than the other agents. The authors supposed that the military command chain might justify, in the participants' minds, the actions of soldiers: a human pilot is seen more blameworthy for cancelling the strike than launching it because self-reliantly terminating the command chain is seen as a moral violation, although no differences in norms postulated to agents were

detected. Thus, if machines are a part of a complex command chain and malfunction or people get killed, somehow the perception of responsibility gets diffused and disappears; although somebody did make the decision to use a machine to achieve this morally relevant task.

However, our own studies conflict with existing literature. Laakasuo et al. (under revisions[4]) studied how people morally judge a hypothetical scenario where either a human nurse or a nursing robot forcefully medicates an unwilling patient. This dilemma juxtaposes two moral principles: the patient's autonomy and the medical establishment's goal to heal the patient. In a series of studies, both qualitative and quantitative (total N > 1300), we found that the people disliked robot-made decisions depending on the type of decision made, and not generally. If the nursing robot decided not to forcefully medicate the patient, the decision was judged similarly than if a human nurse made the same decision. However, forceful medication was only tolerated for a human nurse. These findings are in some tension with findings of Bigman and Gray (2018), wherein mind attribution (or lack thereof) explained the aversion to robots as decision-makers. While people may be generally averse to robots as moral agents, this aversion does not seem to reflect in their judgments about a robot's decisions if those decisions align with what they would prefer a human to decide in a similar situation. Our results also conflict with those of Malle et al. (2019), as it is the robot and not the human agent that is judged more harshly for a specific moral decision.

In a similar vein, Laakasuo et al. (in preparation) presented participants with vignettes (short stories that often depict social events) describing a moral dilemma involving a human or robot coast guard. The guard witnessed a boating accident caused by two intoxicated motorboaters, where three people ended up in water separated by a distance: the motorboaters in one location, and a fisherman in another. The guard had to then decide to either save the

4 Preprint. https://psyarxiv.com/bkhyq/

two motorboaters (utilitarian decision) or the fisherman (favouring the innocent party). The results consistently showed that saving the motorboaters (who caused the accident) was more condemnable than saving the innocent fisherman, but only if the coast guard was a robot. If the guard was a human, both decisions were equally approved.

It seems that robots are held to a "higher" moral standard than humans: people are allowed to choose the "worse" option, but a robot should "know better." Interestingly, the coast guard robot making a utilitarian decision to save the motorboaters was perceived to have "less mind" than a robot deciding to save the fisherman. Utilitarian robots may be seen as cold and calculating, and thus less human. Alternatively, people may consider a robot that (seemingly) takes the moral blameworthiness of the motorboaters into account as more human-like; or people may simply attribute more human-like qualities to robots acting in line with their own morality (not saving the blameworthy motorboaters).Whatever the case, there may be several factors that play into how much human-like thought or feeling people perceive in robots, and in turn how this perceived human-likeness affects judgments on those robots.

As this short review reveals, the *new ontological category* raises its head in situations where machines make decisions about human lives, and humans need to judge whether these decisions are acceptable. With a quick glance, it seems that machines are capable of making near-optimal decisions in multiple domains such as risk management (Lin & Hsu, 2017), medical diagnostics (Elkin et al., 2018), and even in games requiring strategic decision-making (Tegmark, 2017). However, this superficial understanding does not take into consideration that conceptions of "optimal" or "good" might be fuzzy and intuitive rather than sharp and logical. It might be a good idea to delegate moral decisions to machines; it just seems we do not really know how to do that correctly, because our thinking is fuzzy and further complicated by the NOC.

3.2 Transhumanism and Disgust Sensitivity

Previously, we described in Section 2 what transhumanism is as a (normative) philosophy. There have not been many empirical studies on how ordinary people actually feel about transhuman philosophy if it becomes an actuality. Here, we present novel results of work currently under preparation or in press. We describe a number of experiments where we, and others, have studied reactions of ordinary people towards transhuman technologies and technologies that break down the human-machine dichotomy.

Castelo et al. (2019) investigated how individuals who decide to alter their brain functions with either chemicals or brain-implanted chips are perceived as less "human," a phenomenon labelled as *dehumanization*. Dehumanization commonly occurs before intergroup conflicts escalate into full-blown genocides (Arendt, 1951; Haslam, 2006; Haslam et al., 2007). The dehumanized out-group is often described as something less-than-human, animal-like, or less deserving of dignified human treatment. Nonetheless, Castelo et al. (2019) did not attempt to explain *why* dehumanization occurred in their study; or what *motivated* dehumanization of individuals undergoing cognitive or brain enhancement. This question was, however, explored by Koverola et al. (2020b) in a five-study paper (preprint).

The authors investigated whether there were differences in people's reactions when the *memory* or *IQ* enhancement is used to: a) fix an existing ailment, b) achieve optimal human functioning, or c) achieve superhuman functioning. As dependent variables, Koverola et al. 2020 measured: 1) the moral condemnation of the decision to get brain implants, 2) the perceived unfairness of their use, and 3) dehumanization of those individuals deciding to use said cognitive enhancements. The results showed that people were quite accepting of the use

of brain-implant chips unless they were used to gain superhuman abilities. Moreover, the moral foundation of purity (norms about bodily "sanctity": see Part 1 on the Moral Foundations Theory) predicted dehumanization and moral condemnation of memory implants. Further probing revealed that science fiction hobbyism[5] predicted moral approval, and that sexual disgust sensitivity (SDS, See Part 1) was the strongest explaining factor of condemnation and dehumanization of brain-implant chip users (see Figure 1). The authors ruled out competing explanations such as respondents' tendency to oppose new and unknown technologies, medical operations *per se,* or body-envelope violations[6].

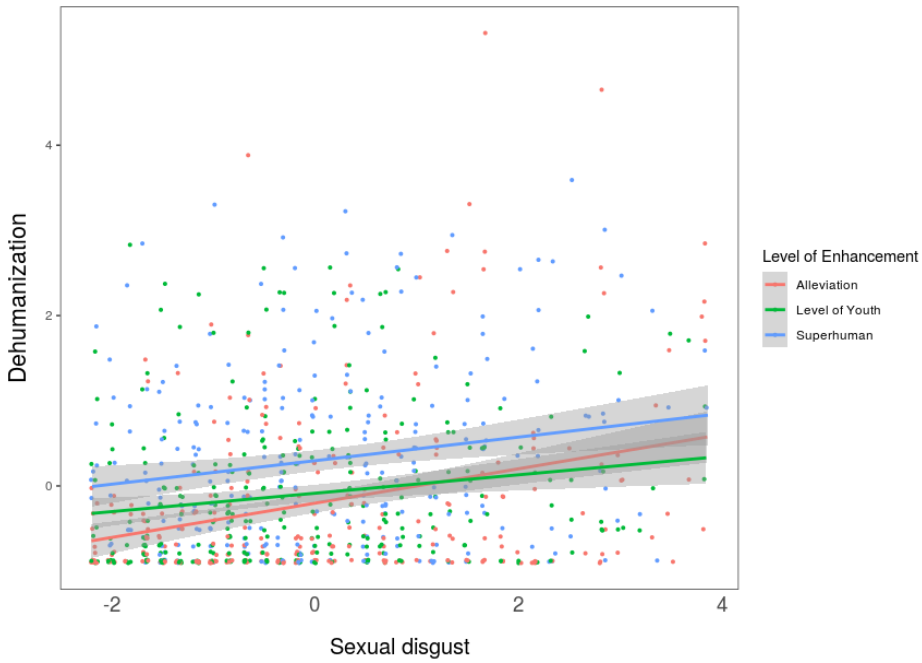

Figure 1. Partial results from an upcoming paper by Koverola et al. (2020b)

In Figure 1, Koverola et al. (2020b) chart results from a series of six experiments where an office worker is suffering from early onset of memory problems and decides to go to the doctor's office for diagnosis. There the patient is given a recommendation of having

5 The science fiction hobbyism scale, used in these studies, measures a general interest and participation in science fiction, with questions about, e.g., participation in conventions, following science fiction series and movies, etc.

6 Body Envelope Violation: invasion of bodily integrity, harm caused to the body that somehow violates its usual status like cuts, injections, fractures, and unwanted penetration.

one of three brain-implanted chips (participants read only one version of the story) with the following potential outcomes: 1) alleviation of the memory problems, 2) return to functioning at the level of youth, or 3) superhuman memory abilities. Participants were asked to rate "how human" the office worker would be after the operation. In this figure, the authors have pooled the data from the studies and we see that: Sexual disgust sensitivity predicts increased dehumanization of brain-implant users (the office worker), irrespective of level of enhancement; however, individuals with superhuman memory capacities are dehumanized more than individuals with normal levels of memory functioning.

The fact that a person's familiarity with science fiction is associated with them having a more positive attitude towards transhuman technology makes intuitive sense. Exposure to new ideas makes them less scary: there are cultural effects to what people judge (alternatively, people drawn to science fiction may share certain personality traits that make them less judgmental in this area). This effect was also observed in another study by Koverola et al. (2020a), where participants judged hypothetical scenarios about robot and human prostitution: the use of both was condemned by participants, but only the judgment of robot prostitution depended on the participants' familiarity with science fiction. Thus, there is some indication that an intuitive effect of familiarity on judgment of very different novel technologies replicates. What is less intuitive is the connection between judgment and sexual disgust.

Why was SDS associated with transhuman technologies? Sexual disgust evolved for mate selection, but has also been co-opted to guard conservative norms; and, apparently, also motivates the condemnation of new technologies. Perhaps our complex modular and categorical cognitive system cannot cope rationally with the blending of the human category with the *new ontological category* (i.e., modern intelligent implant technology). Dehumanization may be triggered by this perceived *mix* of human and machine, which

confuses our biological motivational systems (relating to sexual reproduction). It is hard to imagine a more fundamental blending of ontological categories than that of humans turning into a robot.

Humans have evolved to quickly recognize the difference between the living and the dead, minded and un-minded. But how would humans deal with the ultimate transhumanist dream of uploading one's mind into a supercomputer (Kurzweil, 2012)? This might seem like the most far-off version of science fiction fantasy. However, a copy of a *C. elegans*' (Fessenden, 2014) nervous system has already been placed inside a robot, and a functional copy of part of a rat's brain has been digitized (Markram et al., 2015). In both cases the copy functions similarly to the original. In principle, at least, there is no reason for why this could not be done for the human brain. The movie *Transcendence* (Pfister, 2014) juxtaposes the ethics of self-enhancement, individual freedom, and the conservative public backlash against creating "conscious machines." Clearly, this theme of *mind upload* has deeply enticing moral dimensions for people to produce a multimillion dollar movie.

This theme was recently examined in detail by Laakasuo et al. (2018). In four studies the authors show a familiar pattern previously discussed in the context of brain-implants. Science-fiction hobbyism strongly and independently predicted positive approval of uploading one's consciousness into a computer, whereas sexual disgust and moral purity (independently of each other) strongly and robustly predicted disapproval[7]. The authors also showed that people probably did not consider mind upload as a form of suicide or death, since participants anxious about death and judgmental towards suicide were likely to approve using such technology.

7 Laakasuo et al. (2018) ran several multivariate regression analyses where the associations of independent variables in relation to the dependent variable (moral approval of mind upload) can be investigated while holding the other variables constant. These associations are not causal effects, but they do have predictive value, nonetheless.

One core part of transhumanist arguments for promoting life-extension technologies and cognitive enhancement is the elimination of suffering and increase of life quality. Thus, transhumanism is not that far-removed from the philosophical tradition of utilitarianism. Indeed, many transhumanists are implicitly or explicitly utilitarian in their ethical leanings. They wish to develop transhumanist technologies to promote wellbeing. Tied to this is also the sub-field of AI research known as *AI safety research* (Sotala & Yampolskiy, 2015; Yampolskiy, 2018). In this field, many prominent transhumanists and AI developers analyze risks that humanity might face in seeking to develop human-like Artificial General Intelligence (AGI). One central risk associated with creating an AGI is that it, or its developer, could be a callous, psychopathic, and selfish being with the potential to develop an entity with superhuman capacities.

Focusing more on this issue, we ran several structural equation models on a large online dataset (N = 1000) and found a causal pathway model shown in Figure 2 below. The results once again replicate the statistical effect of sexual disgust on disapproval of mind upload technology. However, one of the fears of the transhumanists seems to be supported, as approval of mind upload technology was linked with Machiavellianism – a personality trait associated with narcissism and psychopathy[8] (see Figure 2; see also Paulhus & Jones, 2015).

8 Note that the definition of Machiavellianism in psychology is simply a collection of specific traits (a person being cold, calculating, etc.). This is different from how the term has been used in, e.g., political theory.

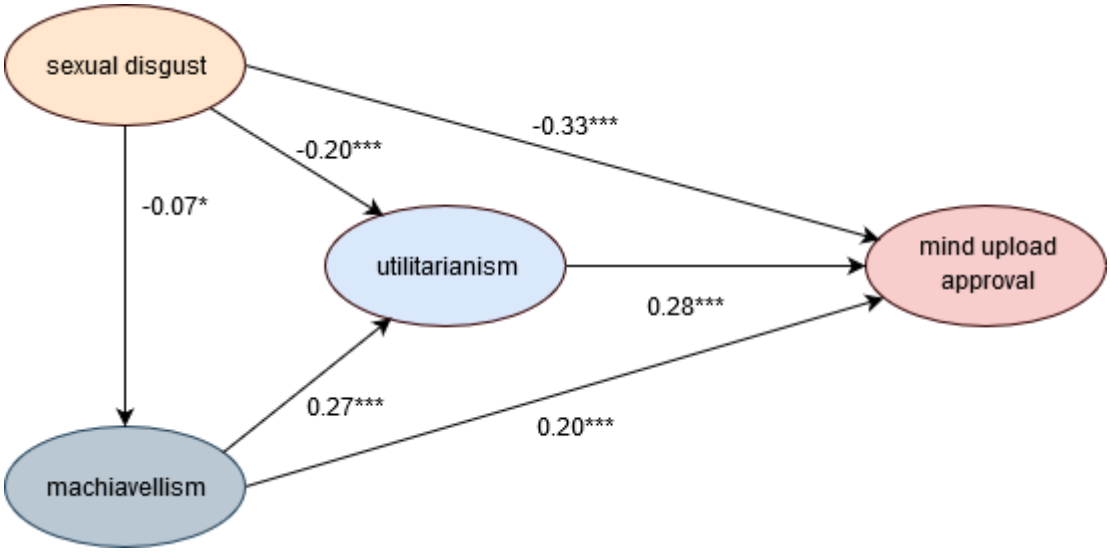

Figure 2. Structural equation model from an upcoming paper by Laakasuo et al. (2020)

In Figure 2, Laakasuo et al. show how Machiavellian tendencies motivate both utilitarian moral choices and approval of mind upload technologies. $N = 1000$; $X^2_{SB(578)}$: 1356,22, CFI= .95, TLI= .94, Robust RMSEA= .039, [.036, .042], SRMR= .043 (indicating an excellent fit between the model and the data). See also Laakasuo et al. (2018).

Machiavellianism is commonly considered as a more functional form of psychopathy, since both dimensions are described as manipulative, callous, and cold (Miller et al., 2017). While Machiavellianism and psychopathy are similar, they are two separate constructs: psychopathy is separated from Machiavellianism mostly by impulsivity and a lack of long term strategizing (Paulhus & Jones, 2015).

This implies that callousness is associated with utilitarian views, which then feed into positive approval of mind upload technology. This study is a good example of how AI safety research[9] can generate hypotheses (in this case, the risk of callous individuals being especially interested in this futuristic technology), which can then be studied verified by moral psychology. From an EP perspective, empathy and sexual disgust explain individuals'

9. AI Safety Research is a specific sub-field of computer science and technology studies that focuses on pre-emptively thinking of strategies of how to avoid pit-falls of creating human-level AIs (e.g. Bostrom, 2014).

interest towards morally volatile future technologies. The issues, as well as the novel results listed here are salient warnings of the challenges posed by the *new ontological category* to our stone-aged moral cognition.

4. FUTURE RESEARCH DIRECTIONS

Crossroads for Transhumanism, AI Ethics, and Moral Psychology

Evolutionary approaches have been effective at generating hypotheses on how humans behave in morally dilemmatic situations. This might not be the case anymore, when the *new ontological category* of AIs and intelligent technologies enter the scene, were they within or outside of our bodies. In the context of the *new ontological category,* there are no obvious immediate hypotheses that could help us understand the implications for our moral cognition. It merely states that new information processing technologies are challenging our moral cognition and gives a plausible explanation for why this is the case. What we specifically need, however, is to gain understanding regarding the deeper process at the level of cognitive structures, as to why we treat robotic decisions and transhuman technologies the way we do. In other words, we need a better/new theoretical framework for hypothesis generation, now that EP approaches are running out of steam.

For instance, the empirical findings presented previously suggest that we should investigate the role of sexual disgust sensitivity in predicting aversion to robots making moral decisions in possibly utilitarian contexts. However, it seems like a bizarre alleyway to go down, since utilitarian robots are not potential disease vectors (i.e., anything, that might carry a pathogen and make us sick), and it is quite difficult to understand why mate choice mechanisms would be associated with the condemnation of robots. Sexual disgust sensitivity specifically seems to be connected to political conservatism (Elad-Strenger et al., 2020), but the associations between sexual disgust sensitivity and moral judgment seems to remain even

after controlling for measures of political views (Laakasuo et al., 2018; 2020). Thus, disgust sensitivity has effects independent of its association with political views (ignoring, for the moment, the question of potential ultimate causes for political views). Theoretically, this connection between disgust and judgment makes little sense, and should be investigated further. One explanation has been offered by Voiklis and Malle (2018): there is no moral cognition as such, and we simply have a collection of social cognitive information processing systems functioning in domains that have been culturally delineated as moral domains. However, this does not tell us *why* a certain set of socio-moral cognitive mechanisms get activated in specific ways with robots and transhuman technologies. Who would have predicted, from evolutionary premises, that mind-uploading is condemned mainly due to sexual disgust mechanisms? This seems obvious post-hoc, but we claim that it would have required exceptional theoretical arguments to produce this hypothesis *a priori*.

Another area that we believe needs more work is in understanding the dynamics of mind perception, dehumanization/infrahumanization, and morality. Will people become more accepting of any and all moral decisions by machines if they are made to resemble humans more? Are “humanized” AVs allowed to make more utilitarian decisions than non-talking AVs, or do we perceive them as something even creepier? Will upbringing among (or even by) increasingly humanized robots disrupt normal cognitive development in children, necessitating countermeasures? What will be the moral status of “enhanced” human beings, as either moral agents or patients? Do perceptions of “enhanced” humans as “less human” also manifest with more clearly visible enhancements such as prostheses, for which there is more cultural history and familiarization? We encourage other researchers to examine these phenomena, since the wellbeing of future generations might be linked to these sorts of mechanisms and their clear understanding.

We also presented the concept of technological animism. This observation, stemming from cross-cultural comparisons (see Part 1) between Eastern and Western cultures implies that we might need to pay more attention to other mechanisms that are associated with mind attribution when we are trying to understand the moralization processes of humans toward robots and transhuman technologies. Previous research implies that human intuitions about souls (not minds) and their purity are also important (Laakasuo et al., 2018; Bering, 2006), but we do not know why. However, it seems clear that human-robot moral interactions are not just about mind (or soul) perception, but also about many auxiliary mechanisms (emotions, perceptions of moral causality, etc.), the role of which needs to be clarified. For example, robots may simply be percieved as something that humans are not as able to negotiate with, as they are with other humans – a potential reason for the increase wariness about the idea of allowing robots to, e.g., make morally complicated medical decisions[10].

As world-changing as the above mentioned technologies are, there are visions of far more disruptive technologies that moral psychology should examine. For example, AI-assisted eugenics would open a space for interesting inquiry where transhumanism, AI, and moral psychological theory intersect. The transhuman idea of increasing the IQ of humanity already has some "prototypes" in existing medical practice. Currently, about 90% of embryos diagnosed with Down's syndrome in prenatal screening are terminated (Morris & Springett, 2014). In some countries, like Iceland, there are almost no individuals with Down's syndrome. In China, genetically modified embryos have already been grown (and widely condemned); police and other state officials have AI technology at their disposal, making it possible to visualize the basic phenotypical characteristics of individuals just from their DNA sample alone (Curtis & Hereward, 2018; Lippert et al., 2017; Schaefer, 2016), and use this as an estimation whether the fetus should be aborted or not.

10 We wish to thank an anonymous reviewer for pointing us to this possibility.

Corresponding technology can theoretically provide means for estimating the characteristics of unborn children, including their IQ, based on the analyzed DNA sequence of the fertilized ova. Within the past 20 years, AI technology and algorithmic data-mining have made it possible to alter the genome of unborn humans (and other organisms) in increasingly reliable ways (Pluysnin et al., 2008; Ritchie et al., 2015). Efficient and automatic computation has quietly made it easier to genetically screen for IQ; essentially making "personal eugenics" a possibility (Regalado, 2018). Given the growing role of machine learning in medicine, AI-based recommendations on embryo screening for "desired" genotypes are a possible, if scary, future development. The idea of screening embryos for intelligence, aided by AI, brings together many of the issues we have discussed: artificial agents affecting human lives, the modification of humanity, even the ultimate EP theme of procreation and survival. As with many other themes discussed here, we do not really know how people feel about these technologies. What moral cognitive processes are activated, if culture shifts towards accepting the use of these technologies (Rozin, 1999)? We encourage researchers to study people's responses to seemingly far-fetched or "sci-fi" ideas. We are not in a technological utopia or dystopia, nor likely to get there very soon, but some of the themes dealt with in the moral psychology of robotics or transhumanism are getting closer to real-world relevance, or are already there. Self-driving cars, robot companions to children, and at least pharmacological cognitive enhancement are already happening (Thompson, 2014). Our suggestion to the problem of being out-paced by technology is to ask questions about technologies that are not here (yet).

5. CONCLUSION

To have a basic understanding of the problems that artificial moral agents and transhumanistic technologies pose, we should (or must) use the combined tools of philosophy, evolutionary psychology, moral psychology, technology studies, and

anthropology. We hope to have shown that there are indeed some previously unencountered questions and problems with which our evolved moral cognition must deal. We have shown how evolutionarily old cognitive mechanisms (e.g., sexual disgust sensitivity) are unexpectedly linked with transhumanist technologies. We have also shown that although robots are not "minded" in the same way as biological beings are, people still treat them as having at least a degree of a mind. Clearly robots are not treated like rocks, trees, other tools, or even like "mere machines," but neither are they treated as humans or animals (see Part 1).

Moral dilemmas involving self-driving cars, killer drones, nursing robots, and rescue robots, when analyzed from this novel point of view (combining as mentioned, philosophy, evolutionary psychology, moral psychology, technology studies, and anthropology) reveal new vantage-points into our own moral cognition and its functioning. We do treat them as if they have minds, but to which degree seems to depend on the type of decisions (e.g., utilitarian vs. deontological) these devices make. We also recognize that many of the findings, models, and theories presented here might only apply to Western cultures. However, research by Awad et al. (2018) suggests there are some cross-cultural universals further highlighting the need for bio-cultural approaches that take into consideration both evolutionary modular models and cultural influences.

We have discussed how the existing moral psychological theories help us to understand our own reactions when robots make moral decisions or merge with the human brain. What is the evolutionary explanation, or even a hypothesis, for why sexual disgust sensitivity predicts moral approval or condemnation of mind upload, or brain implants? The *new ontological category* is a cybernetic cosmic trickster monkey throwing its wrench into our moral cognitive system. The ensuing mess is unique in the history of humanity and should be studied.

References

Arendt, H. (1951). *The origins of totalitarianism*. Schocken Books

.

Bering, J. M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences*, *29*(5), 453–462. https://doi.org/10.1017/S0140525X06009101

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Blackford, R. (2003). Who's Afraid of the Brave New World? *Quadrant*, *47*(5), 9.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Bostrom, N. (2005). A History of Transhumanist Thought. *Journal of Evolution and Technology*, *14*(1).

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Burdett, M. (2014). The Religion of Technology: Transhumanism and the myth of progress. In C. Marcer & T. J. Trothen (Eds.), *Religion and transhumanism: The unknown future of human enhancement* (pp. 131–147). ABC-CLIO, LLC.

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

Children. United Nations Global Issues. https://www.un.org/en/sections/issues-depth/children/

Curtis, C., & Hereward, J. (2018, May 4). How Accurately Can Scientists Reconstruct A Person's Face From DNA? *Smithsonian Magazine*. https://www.smithsonianmag.com/innovation/how-accurately-can-scientists-reconstruct-persons-face-from-dna-180968951/

Elad-Strenger, J., Proch, J., & Kessler, T. (2020). Is Disgust a "Conservative" Emotion? Personality and Social Psychology Bulletin, 46(6), 896–912. https://doi.org/10.1177/0146167219880191

Elkin, P., Schlegel, D., Anderson, M., Komm, J., Ficheur, G., & Bisson, L. (2018). Artificial Intelligence: Bayesian versus Heuristic Method for Diagnostic Decision Support. *Applied Clinical Informatics*, *09*(02), 432–439. https://doi.org/10.1055/s-0038-1656547

Fessenden, M. (2014, November 19). We've Put a Worm's Mind in a Lego Robot's Body. *Smithsonian Magazine*.

Frischmann, B., & Selinger, E. (2018). *Re-Engineering humanity*. Cambridge University Press.

Guillette, S. (2019, December 6). Your new lifeguard may be a robot. Verizon. https://www.verizon.com/about/our-company/fourth-industrial-revolution/your-new-lifeguard-may-be-robot

Haslam, N. (2006). Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Haslam, N., Loughnan, S., Reynolds, C., & Wilson, S. (2007). Dehumanization: A New Perspective. *Social and Personality Psychology Compass*, *1*(1), 409–422. https://doi.org/10.1111/j.1751-9004.2007.00030.x

Huxley, J. (1957). Transhumanism. In *New Bottles for New Wine*. Chatto & Windus.

Kass, L. R. (2003). Ageless Bodies, Happy Souls: Biotechnology and the Pursuit of Perfection. *The New Atlantis: A Journal of Technology & Society*, *1*, 9–28.

Koverola, M., Drosinou, M., Palomäki, J., Halonen, J., Kunnari, A., Repo, M., Lehtonen, N., & Laakasuo, M. (2020a). 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*, *11*(1), 233-249. doi: https://doi.org/10.1515/pjbr-2020-0012

Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I., Sundvall, J., & Laakasuo, M. (2020b, June 30). NON-HUMAN SUPERHUMANS - Moral Psychology of Brain Implants: Exploring the role of situational factors, science fiction exposure, individual differences and perceived norms. https://doi.org/10.31234/osf.io/qgz9c

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. Viking Penguing.

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

Lin, P., Mehlman, M., Abney, K., & Galliott, J. (2014). Super Soldiers (Part 1); What is Military Human Enhancement. In S. J. Thompson (Ed.). *Global Issues and Ethical*

Considerations in Human Enhancement Technologies (pp. 139–160). IGI Global. Hershey: PA. https://doi.org/10.4018/978-1-4666-6010-6.ch008

Lin, P., Mehlman, M., Abney, K., French, S., Vallor, S., Galliott, J., Burnam-Fink, M., LaCroix, A. R., & Schuknecht, S. (2014). Super Soldiers (Part 2): The Ethical, Legal, and Operational Implications. In S. J. Thompson (Ed.). *Global Issues and Ethical Considerations in Human Enhancement Technologies* (pp. 139–160). IGI Global. Hershey: PA. https://doi.org/10.4018/978-1-4666-6010-6.ch008

Lin, S.-J., & Hsu, M.-F. (2017). Incorporated risk metrics and hybrid AI techniques for risk management. *Neural Computing and Applications*, *28*(11), 3477–3489. https://doi.org/10.1007/s00521-016-2253-4

Lippert, C., Sabatini, R., Maher, M. C., Kang, E. Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., Yocum, K., Wong, T., Zhu, M., Yang, W.-Y., Chang, C., Lu, T., Lee, C. W. H., Hicks, B., Ramakrishnan, S., … Venter, J. C. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences*, *114*(38), 10166–10171. https://doi.org/10.1073/pnas.1711125114

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. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (Vol. 95, pp. 111–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., Kahou, G. A. A., Berger, T. K., Bilgili, A., Buncic, N., Chalimourda, A., Chindemi, G., Courcol, J.-D., Delalondre, F., Delattre, V., … Schürmann, F. (2015). Reconstruction and Simulation of Neocortical Microcircuitry. *Cell*, *163*(2), 456–492. https://doi.org/10.1016/j.cell.2015.09.029

Melson, G. F., Kahn, P. H., Beck, A., Friedman, B., Roberts, T., Garrett, E., & Gill, B. T. (2009). Children's behavior toward and understanding of robotic and living dogs. *Journal of Applied Developmental Psychology*, *30*(2), 92–102. https://doi.org/10.1016/j.appdev.2008.10.011

Miller, G. (2013). Chinese Eugenics. *Edge.Org*. https://www.edge.org/response-detail/23838

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

More, M. (2010). The Overhuman in the Transhuman. *Journal of Evolution and Technology*, *21*(1), 1–4.

Morris, J. K., & Springett, A. (2014). *The National Down Syndrome Cytogenetic Register*.

O'Connell, M. (2017). *To be a machine: Adventures among cyborgs, utopians, hackers, and the futurists solving the modest problem of death*. Granta Publications.

Paulhus, D. L., & Jones, D. N. (2015). Chapter 20—Measures of Dark Personalities. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (pp. 562–594). Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00020-6

Pfister, W. (2014, April 10). *Transcendence*. Warner Bros. Pictures.

Plyusnin, I., Evans, A. R., Karme, A., Gionis, A., & Jernvall, J. (2008). Automated 3D Phenotype Analysis Using Data Mining. *PLoS ONE*, *3*(3), e1742. https://doi.org/10.1371/journal.pone.0001742

Regalado, A. (2018, April 2). DNA tests for IQ are coming, but it might not be smart to take one. *Technology Review*. https://www.technologyreview.com/s/610339/dna-tests-for-iq-are-coming-but-it-might-not-be-smart-to-take-one/

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85–97. https://doi.org/10.1038/nrg3868

Rozin, P. (1999). The Process of Moralization. *Psychological Science*, *10*(3), 218–221. https://doi.org/10.1111/1467-9280.00139

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

Sotala, K., & Gloor, L. (2017). Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. *Informatica*, *41*, 501–505.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Thompson, S. J. (Ed.). (2014). *Global issues and ethical considerations in human enhancement technologies*. IGI Global. Hershey: PA. https://doi.org/10.4018/978-1-4666-6010-6

Voiklis, J., & Malle, B. F. (2018). Moral cognition and its basis in social cognition and social regulation. In K. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 108–120).

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the

Dead. *Psychological Science*, *24*(8), 1437–1445. https://doi.org/10.1177/0956797612472343

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Yampolskiy, R. V. (2018). *Artificial intelligence safety and security*. Chapman and Hall/CRC.