þÿSocio cognitive biases in folk AI ethics and risk

Laakasuo, Michael

2021-11

Laakasuo , M , Herzon , V , Perander , S , Drosinou , M-A , Sundvall , J , Palomäki , J P & þÿVisala , A 2021 , ' Socio cognitive biases in folk AI ethics and risk dis , no. 1/2021 , pp. 593-610 . https://doi.org/10.1007/s43681-021-00060-5

http://hdl.handle.net/10138/339313
https://doi.org/10.1007/s43681-021-00060-5

AI and Ethics (2021) 1:593–610
https://doi.org/10.1007/s43681-021-00060-5

REVIEW

Socio-cognitive biases in folk AI ethics and risk discourse

Michael Laakasuo[1] · Volo Herzon[2] · Silva Perander[3] · Marianna Drosinou[1,2] · Jukka Sundvall[1] · Jussi Palomäki[1] · Aku Visala[4]

Received: 27 January 2021 / Accepted: 23 April 2021 / Published online: 5 June 2021

Abstract
The ongoing conversation on AI ethics and politics is in full swing and has spread to the general public. Rather than contributing by engaging with the issues and views discussed, we want to step back and comment on the widening conversation itself. We consider evolved human cognitive tendencies and biases, and how they frame and hinder the conversation on AI ethics. Primarily, we describe our innate human capacities known as folk theories and how we apply them to phenomena of different implicit categories. Through examples and empirical findings, we show that such tendencies specifically affect the key issues discussed in AI ethics. The central claim is that much of our mostly opaque intuitive thinking has not evolved to match the nature of AI, and this causes problems in democratizing AI ethics and politics. Developing awareness of how our intuitive thinking affects our more explicit views will add to the quality of the conversation.

Keywords Artificial intelligence · Moral psychology · Moral philosophy · Machine ethics · Cognitive bias · Folk theories · Categorization

1 Introduction

Our everyday thinking, in dealing with the world around us, mostly relies on evolved cognitive classifications and categorizations (see Atran [5], Boyer [23]). Due to these evolved capacities, we are able to predict changes in our environment and update these predictions rapidly [118]. In this paper, we first describe the cognitive process of categorizing and then show why the concept of artificial intelligence (AI) does not easily fit into our intuitive everyday categories. This lack of fit means that AI can be viewed as a moderately counterintuitive concept [116]. The fallacies and biases encountered in discussions of AI ethics among the general public are, we argue here, partially explained by how AI is a counterintuitive concept [104, 105].

AI commonly refers to technology that has the capacity for making (mechanical) decisions either autonomously or through enhancing decisions made by humans. Thus, the concept of AI covers a wide range of technical programming and statistical algorithmic solutions to typically local and narrowly defined problems, rather than referring just to a deep learning neural network [127]. It is hard to define AI precisely because it is hard to define intelligence precisely [127, 179]. For the purpose of this paper, we define intelligence as goal-oriented and purposeful action taken in an at least partially predictable environment [111]. We will refer to AI in its narrow sense as an algorithm that functions purposefully in an at least partially predictable environment [179].[1]

Jussi Palomäki, Aku Visala are equal contributions.

✉ Volo Herzon
volo.herzon@gmail.com

[1] Department of Digital Humanities, Cognitive Science, Faculty of Arts, University of Helsinki, Helsinki, Finland

[2] Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[3] Department of Computer Science, Faculty of Science, University of Helsinki, Helsinki, Finland

[4] Department of Theology and Religious Studies, University of Helsinki, Helsinki, Finland

[1] We refer both to academic sources and other recent reports, news pieces, and commentaries. Academic sources and books are referenced traditionally and the other sources in footnotes. This helps keep the theoretical discussion separate from mundane examples.

2 Anthropomorphizing across categories

Folk theories are an innate human tendency of predicting and explaining various natural phenomena without relying on scientific education (see Guglielmo et al. [65]), and they heavily rely on our capacity to intuitively categorize our perceptions. Scientific theorizing and experimental work depend on our capacity for reflective thinking, and do not fully replace our intuitive thinking [156]. This is most apparent in circumstances where we desire knowledge but have little science to rely on [156]. More specifically: folk psychology aims to predict the behaviour, emotions, and cognitions of other humans (or other animals); folk biology, in turn, aims to predict the developmental history and the reproductive mechanisms of various organisms; while folk physics aims to predict and explain the forms and motions of non-living objects and substances. Our capacity to form intuitive explanations has emerged through selection pressures in our evolutionary history. Our ancestors developed everyday, non-scientific, explanations of various natural phenomena, such as earthquakes, the seasons, and the motions of celestial bodies [24]. For the folk theories to function, the human mind must be capable of classifying phenomena into various categories. Each category invites the application of different, mostly automatic cognitive inference rules. Such categories include tools [147], predators [24], pets [4], and plants [6]. In shorter time spans, humans can create new categories through cultural processes; however, these cultural categories do not necessarily function equivalently to our more basic categories [104, 105].

One of the key categorical distinctions is between agents (such as humans and some animals) and non-agents (plants, artifacts, tools, etc.; Barrett [12]). The capacity to make this distinction has been crucial for the survival of humankind. Being able to discriminate between friendly and unfriendly agents with only a sparse amount of input cues from the environment could provide a survival advantage for any organism [109]. Agents are typically conscious biological beings whose behavior can be explained by folk psychology, while non-agents, on the other hand, require folk physics. Recent research in developmental psychology has shown that children can distinguish agents from human-made artifacts at a very early age [87]. Similarly, at an early age children comprehend the functions of different kinds of tools and other artifacts [91].

We argue that AI poses a significant problem to our categorization capacities. We have no evolutionarily developed innate capacity for recognizing complex information-processing systems, such as AIs and robots, which simultaneously appear to be both artifacts and artificial agents. Artificial agents are entities which resemble natural agents but lack both selfhood and, likely—and as a background assumption in this article—consciousness. On an evolutionary scale, there simply has been no time for human cognition to develop a natural capacity to comprehend autonomous technology and predict its behavior. Indeed, some scholars in robotics and developmental psychology now use the neologism new ontological category (NOC), in reference to AIs and robots, to explain the origins of at least some of these problems [88].[2] There is no precedent in the history of our planet that an object made from lifeless material starts moving, behaving, and acting as if it were a living creature.

Of course, there have long been machines designed to imitate human or animal behavior, but these automata have always been extremely primitive compared to contemporary and developing technological systems [17]. Currently, however, many AIs make independent ethical decisions which immediately or indirectly affect human wellbeing [177], even if the decisions are mechanical and made without consciousness (see [42]). Our cognitive capacity for folk theories developed in the Upper Pleistocene environment (circa 2 million–200,000 years ago, Tooby and Cosmides [171]). Because, at that time, there were no robots, computers, formal algorithms or cybernetic systems to interact with, we lack the natural capacity to categorize them and predict their (non-conscious, logical and probabilistic) behavior based on evolved folk theories alone. Consequently, we need to resolve the lack of innate cognitive categories through developing sufficiently precise cultural categories [105].

Robots and AIs challenge our stone age minds in deeper ways, too, and our folk theories struggle to make sense of them. Our evolved automatic cognitive systems can be trained to process novel domains (such as our face-recognizing system also specializing in recognizing birds if we become ornithologists), but there is always a loss of fluency compared to use within the original domain [24]. We categorize robots and AIs, depending on their surface appearance, inconsistently as animals, tools, toys, or children, while they are none of these [25, 35]. For instance, witnessing a robot dog being kicked may make us frown and feel some form of compassion like we would towards a real dog [123, 124]. Similarly, robots designed for social interaction, such as the Paro seal used in elderly care, activate our positive social emotions [122]. These reactions are partially explained by the phenomenon of anthropomorphizing, the tendency to think and talk of non-human and non-living objects as if they have feelings, desires, personalities. A phenomenon

[2] To avoid confusion, it is important to note that by *ontological categories* we do not refer to any classification scheme normalized by any specific culture. Rather, we refer to the implicit information processing predispositions of human cognition which, in part, shape all cultures.

that—as we know in light of developmental psychology—is also apparent in children when they project humanlike mental capacities to soft toys and other objects [60].

Here, we approach anthropomorphizing as a cognitive and biological phenomenon universally shared by all humans without inspecting its cultural variability. Rather than judging the potential value of anthropomorphizing, we describe some possible folk ethical challenges stemming from it.

Anthropomorphizing and the related oversensitivity of agency-detector systems have notably been studied within the cognitive science of religion. These systems are hypothesized to be at least partially responsible for our tendency to sometimes perceive agency where there is none, and also seem to partially explain the popularity of beliefs in supernatural agents (such as gods, ancestors, and spirits) [13]. Religions, like other cultural phenomena from folklore to militaries, are seen as by-products of different cognitive systems (theory of mind, contamination avoidance, kinship recognition, linguistic competence, etc.). Be it anthropomorphism [109], action representation [66] or episodic and autobiographical memory [108], these capacities now work beyond their original functioning and domains. Robots and AI systems commonly activate these same mind perception mechanisms [168], but our anthropomorphizing tendencies lead us astray [178]. Problems arise when robots and other AIs are specifically projected to have humanlike or partially humanlike minds, and are regarded as sentient, intelligent, and feeling beings. Indeed, the actions of robots that look like humans are judged differently from those that do not [101]. However, such superficial assessments are unwarranted, since robots and AIs function according to different principles from those of the human mind. Their cognitive functioning is based on algorithms and probabilistic computation, both of which are particularly challenging to grasp for our folk cognitions [38, 68, 131, 151].

We humans are generally not very good at logical and axiomatic reasoning. We have trouble comprehending conditional statements like syllogisms [47, 94]. Without extensive training, perceiving and estimating mathematical probabilities is counterintuitive and hard [151], a fact well recognized in evolutionary psychology. A classic example is the gambler's fallacy. We tend to intuitively expect the probability of one event to be related to the outcome of another, even when the events are completely independent (e.g., after a coin is flipped twice and lands on heads both times, it is expected to more likely land on tails with the next flip; Rabin and Vayanos [148]). Moreover, computer programming is generally very difficult and time consuming for people to learn [136]. Programming and understanding software rely on algorithms which are complex concatenations of conditional statements. Current machine learning algorithms are, in turn, based on the concept of probability. Thus, understanding both conditional statements and probability is crucial for understanding how AI functions overall and specifically in ethical contexts. We propose that because people are not capable of comprehending algorithms and probabilities without extensive education, they cannot be expected to accurately comprehend AI technologies or ethical problems related to them.

Given enough time, we might culturally develop a category adequately guiding us in relation to and resembling the new ontological category (which contains robots and AIs). By doing so we could avoid many of the problems raised here. This category would build on features from the existing cognitive categories, such as animals, artifacts, and humans [125]. This cognitive development could be achieved through widespread and systematic education, accommodated by a cultural shift. One starting point would be to increase such education in the elementary and high school curricula. If this project were successful, people would learn to adequately understand the functional principles and behavior of AIs, and regulate their own intuitive reactions towards them.

3 Is it ok to abuse a cute robot?

The problems of anthropomorphizing and miscategorization become readily apparent when we examine social robotics. Social robots have some level of non-conscious comprehension of human social interaction dynamics and norms, and such robots are capable of behaving and communicating with humans in various situations. Social robots, such as the previously mentioned Paro seal, are usually purposefully designed to appeal to human emotions and evoke empathy. For instance, cute and large-eyed care robots remind us of children and baby animals [97]: they appear innocent, helpless, and difficult to ignore. In such cases, robot designers purposefully amplify our anthropomorphizing tendencies and appeal to our universal propensities for compassion towards the innocent and vulnerable.

It has been a matter of debate in the ethics of AI as to what extent it is appropriate to evoke emotions, such as compassion or promote anthropomorphizing attitudes [119, 158]. Is it deceitful to nudge a person to feel compassion towards a being which is incapable of reciprocating, let alone experientially understanding compassion?

Sometimes, relationships resembling authentic social bonds are formed with care robots. People tell their care robots stories, reminisce on the past, stroke and pet them, and cry in their presence [157]. Deceit is a significant moral risk especially when those deceived are people whose cognitive capacities have been diminished, for example, in seniors with dementia [157, 176], because the machines are specifically designed in a way that elicits pro-social reactions.

The use of care robots in elderly care risks leading to the *infantilization* of the elderly: the treatment of the elderly like little children playing with their toys rather than adults with autonomy in many matters [157]. This is how seemingly well-meant manipulation of human social cognition is revealed to be morally problematic and even dangerous. There is evidence that experiencing positive emotions and sharing negative emotions can promote wellbeing [122], but we do not have a clear picture of what the long-term consequences of experiencing emotions based on "deceit" are. It remains worth considering how we could bring about positive emotions in elderly care, without relying on illusion and deceit.

Another type of social interaction in which we encounter miscategorization is that of callous treatment. How should we relate towards robots being damaged or "abused" [178] ? Due to our anthropomorphizing tendencies, some actions towards a robot are interpreted as abuse and thus immoral, even when the action caused no harm, that is experienced suffering or thwarted desires (ibid.). Boston Dynamics, a US robotics company (www.bostondynamics.com) produces robots that move in ways resembling the movement of humans and other animals. When such robots are pushed or kicked, many people readily interpret these actions through our folk psychology lens as "bullying", which in turn causes empathy towards the robots and negative emotions towards the "bully" [123, 166, 183].

Researchers studying robot "abuse" have suggested that rather than judging the action itself, people judge the "abuser" for being callous and lacking compassion [29, 30]. Conversely, banging a computer keyboard or tossing a mobile phone in anger is usually not regarded as a callous act [30, 149]. Such cases reveal how a robot is intuitively categorized as some kind of quasi-human agent. Comparably, the killing of virtual characters in a computer game usually does not evoke similar negative emotions as the "abuse" of robots. Ultimately, in virtue of how they process information mobile phones, computer game virtual characters, and human-like or cute robots, are all equal.

Even if it were in some way irrational to worry about the wellbeing of a robot toy, we think it is worthwhile to consider the impact on the people who interact with the robots [36]. Indeed, some philosophers (following Kant) have proposed an analogous case in how the treatment of animals affects us. In their view, even if animals lack self-awareness and are incapable of suffering, treating them in a cruel way would be harmful to us humans [4, 36, 86]. The explicitly indifferent attitude towards the suffering of "senseless animals" could also contribute to our disposition to treat other humans callously. This claim could be defended by invoking the need of moral education. According to this view the capacities needed for moral action, such as moral imagination, empathy, and compassion, may be to some extent innate. However, the argument goes, they nevertheless require practice and training to function well [114]. Analogously, the way we treat robots could affect us. This view becomes plausible when we consider how our intuitive categorization and explicit views on whether a being is a moral patient (i.e., worthy of moral consideration) do not necessarily coincide. Empirically, this view finds some support by the finding that routine killing of animals by slaughterhouse workers predicts lower levels of well-being for the workers themselves, compared to other "dirty" jobs, such as cleaning and elderly care [10]. One hypothesis for this effect is our tendency to feel empathy towards other species in pain (ibid.), even despite consciously adhering to the belief that animals are non-conscious. Another mechanism might be the toll of the work on the empathetic capacity. Analogously, due to the anthropomorphizing of robots, in our interactions with them, our intuitive moral capacities might be employed and thus, implicitly affected even while we reflectively acknowledge that the robots are not moral patients. This mechanism would function regardless of the robot's actual moral status [114].

The same phenomenon can become a moral psychological problem during the era of AIs and robots [54]. When our everyday reality is populated by various intelligent systems which lack the status of moral patiency, people might become accustomed towards cruelty and indifference. Because we sometimes think of robots as if they were alive and conscious, we may implicitly adopt patterns of behavior that could negatively affect our relationships with other people. In a society which welcomes social robots, such as care robots and sex dolls (see [95]), a person's moral development might be influenced by how they themselves and others around them treat those robots [166]. Due to these moral psychological reasons, we find merit in educating people to approach robots as moral patients even while acknowledging that they are unlikely actual moral patients. We would even go so far as to expect robots to be treated with respect [142].

4 Anthropomorphism and the ideal of rationality of artificial intelligences

Artificial intelligence may in some circumstances muddle our sense of responsibility. Ethical blindness refers to a lowered sense or unawareness of immediately present ethical problems, and it can also happen in social and digi-social contexts [141]. More specifically, an AI may be anthropomorphized as an authority and perceived responsible for decisions made by people [1, 37]. An example of ethical blindness is the case of United Airlines in 2017 when a passenger was forcefully removed from an overbooked flight.[3] There were no volunteers to give up their seat, so

[3] https://www.nytimes.com/2017/04/10/business/united-flight-passenger-dragged.html.

the decision on who was going to be removed was made by a software algorithm. During the event, no staff member questioned the decision or was willing to deviate from it, which led to the situation ending with the passenger being violently removed from the aircraft; the passenger suffered a concussion and lost a tooth. After the incident, the market value of United Airlines fell by hundreds of millions of dollars, and the airline eventually ended up paying the passenger a considerable compensation.[4] In hindsight and as the general rule, the recommendation made by the algorithm should be open to being questioned, challenged and carefully reflected upon.

In the situation described above, the "emotionally cold" system was perceived as a reasonable authority responsible for making the decision. Moreover, the system might have been anthropomorphized in a very specific way: either explicitly or intuitively, the system was attributed with unbiased rationality. The staff found it implicitly acceptable to regard the system's recommendation as reasonable and binding. However, in reality, the system's decision was based merely on rules according to which it had been programmed or trained to act. These rules did not take into account crucial considerations in the context of social interaction, such as the possibility of emotional escalation, the need for diplomacy and creative *ex tempore* solutions. It remains a topic for further research to establish to what extent the role of AI systems affects the decisions in such cases above inflexible corporate hierarchies or other causes unrelated to AI. This is a general problem associated with rigid organizations [50], but we suspect that the use of AI technology will likely amplify this problem.

The key issue in relying on the assumed ideal rationality and infallibility of AI systems is that the functional principles of these systems are, for the most part, opaque [32]. Even now, some AI systems, and especially machine learning algorithms, are so complex that our cognitive apparatuses are incapable of comprehending their functionality and logic [11, 127, 179]. Within ethics, this is known as the black box problem [81]. A machine learning algorithm can be trained to perform some task well, but it is practically impossible for a human to follow the actual decision process behind its performance. There is ongoing ethical discussion on this problem, especially concerning the transparency of algorithmic decisions affecting human lives [107], but we are not going to address it here.

Of interest in this context is the psychological side to the black box problem. We have the tendency to view the decisions made by a black box more reliable than those made by humans, since algorithms are often perceived as coolly rational or even perfectly rational [110]. At this stage of AI development, it is extremely unlikely that AI systems could process all the sources of tacit information on which human behavior constantly relies upon [179]. For example, if a member of the flight staff would have tried to persuade another passenger to give up their booking for a price of several thousands of dollars (the originally offered price was 800$), the airline might have avoided the subsequent PR catastrophe and the loss of millions of dollars. The airline was incapable of considering this possibility, and the staff was not willing to or did not dare question the decision of the airline's representative, the AI.

5 The "they only do what they have been programmed to do" fallacy

There are further problems caused by the lack of fit between our basic cognitive categories and the principles underlying machine learning and other types of AI. We stated above that AI does not neatly fall into categories, such as "animal", "artifact" or "agent". Yet, these categories guide our intuitive thinking, so we will look at them more closely. We want to draw attention to what happens if AI is viewed solely through the categories of "artifact" and "tool". It is common to think that AI, or more broadly, any computer executing programs, is doing simply what it has been programmed to do. This assumption is misleading and affects the discussion about AI ethics considerably [110]. It is different to explicitly program an AI to perform a task than to program it to autonomously learn to perform the task. In both cases, the AI indeed only does what it has been programmed to do, but especially in the latter case it is hard for humans to intuitively follow and predict the complex, data-driven and probabilistic decision making; hence, the aforementioned black box problem and the surprise about the way a learning system will perform a given task [106]. "They only do what they have been programmed to do" is true, but falsely implies that we can always tell what they will do.

For example, many reinforcement learning algorithms can learn by creating their own subgoals [167, 184]. In other words, they are not simply executing subgoals determined by their programmers, but rather discover them independently [139, 167]. In 2013, the researchers at DeepMind Technologies developed an AI, based on a reinforcement learning algorithm, which learnt to play several Atari video games and even outcompeted the best human players in some of the games [128, 128, 167]. What is remarkable is that the only input the algorithm received was what a human player would see from the screen: a set of changing pixels. Based on this data, the algorithm worked to maximize its score in many subsequent playthroughs and over time through trial and error, learned to play the game. The AI learned to do exactly what contributed to a maximum score in each

[4] https://www.nytimes.com/2017/04/11/business/united-airline-passenger-overbooked-flights.html.

individual game. It made no difference whether the game was about flying an airplane or killing virtual characters.

AIs that are based on reinforcement-learning are capable of finding unexpected solutions to well defined problems. They are, ultimately, (simple) cognitive artificial agents learning complex behavioral patterns in different environments. For example, for an AI to kill virtual characters, there is no need to explicitly program an AI to kill, if killing counts towards maximizing the score. The system needs only to be programmed to maximize its score, after which it can quickly become extremely skilled in any specific task. An AI created by the OpenAI project learned to play the popular computer game *Dota 2*. It managed to develop complex predatory ambushes and feint strategies to maximize its score.[5] Such AIs are also capable of unpredictable behavior, as they can learn patterns or strategies unknown to humans. An example of this is the *AlphaGo* algorithm, which used completely novel strategies in the game Go [167]. Another example is *AlphaStar* in the popular real time strategy game StarCraft 2. One professional player called AlphaStar's playing style "unimaginably unusual…[and that it] makes you question how much of StarCraft's diverse possibilities pro players have explored" [164]. These kinds of AI agents could theoretically be trained to kill people in extremely realistic war simulations. Autonomous vacuum cleaners and unmanned aerial vehicles are already tested and trained in virtual environments (e.g., [44]), and when the test results are sufficient, the code is easily copied to an actual physical machine.

Recent history already offers examples on how difficult it can be for humans to predict algorithms' behavior. To illustrate the issue, let us consider stock trading algorithms. They are usually reliable in trading stocks in a complex environment with many human agents [140]. However, when many algorithms are brought to the stock market and these algorithms have not been tested against both humans and other algorithms, unexpected feedback loops might arise. These can cause temporary stock crashes, and indeed, at least three such cases have already been reported.[6] Because we have a tendency to overlook such problems and lack intuition on what it actually means for the AI to "only do what it has been programmed to do", we argue that it would be good to pay institutional and legal attention to algorithm testing and evaluation in different environments. However, unfortunately even that is no guarantee that an AI will work safely and as expected, if novel elements are introduced in its environment [78].[7]

The issue can be presented in the following way: reinforcement-learning-based algorithms are programmed to *learn* (for example, to behave in a way which maximizes a game score). In this sense, they do exactly what they are programmed to do, but what they learn is also determined by the separate and unprogrammed environment in which the algorithm is used. Even tiny unplanned variations in the environment can potentially lead to the AI learning different things, which results in behavior that has not been explicitly programmed by anyone. One of the more striking examples of this is the phenomenon named adversarial attack. In adversarial attacks image recognition algorithms (or recognition algorithms for other kinds of input data) are completely fooled when random noise with specific parameters (in frequency or shape), invisible to human eyes, is injected into photographs [127]. This results in the algorithms categorizing various objects as something entirely different. For an example, an algorithm might classify pictures to show ostriches, even if, for us, the pictures clearly and unambiguously show dogs, buildings, etc. [134, 154].

The reason why learning algorithms are useful is the same as the reason why they are problematic from an ethical perspective: they produce novel and unexpected solutions. Their learning capacity is what makes it so difficult to predict the actions of an AI [78]. Even carefully tested algorithms can act in unpredicted ways in novel environments beyond where they were tested in. It is thus difficult to assess in advance in which environments the AI would function as planned and which its actions would be unexpected and undesired. AIs lack the natural and unspecific moral chokes and restraints, which at best prevent humans from making morally catastrophic choices. However, it makes no difference for an AI-guided robot whether it kills virtual people in a computer game or real humans in a war. It is not necessary for an AI to "desire" anyone's death for it to be an efficient action towards its set goal.[8]

[5] https://www.wired.com/story/can-bots-outwit-humans-in-one-of-the-biggest-esports-games/.

[6] https://www.sec.gov/news/studies/2010/marketevents-report.pdf; https://www.bbc.com/news/business-42959755; https://www.theguardian.com/business/2013/apr/23/ap-tweet-hack-wall-street-freefall.

[7] https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-agents-startle-researchers-with-unexpected-strategies-in-hideandseek; https://www.bbc.com/future/article/20170410-how-to-fool-artificial-intelligence.

[8] Cognitive scientist Steven Pinker addresses risks related to AIs by pointing out the fallacy of assuming AIs to have motivations towards ends harmful to humans [145]. Given that no such motivations exist, he continues that humans would not be capable of programming a powerful AI and at the same time setting it off towards far reaching obviously harmful goals; and if the AI did show unintended consequences, it could simply be turned off (ibid.). We agree with Pinker to the extent human institutions are willing and capable of averting conflict and socially adverse outcomes.

6 Artificial intelligence: morally relevant, even if non-conscious

We described above what kind of problems stem from categorizing AI as primarily an "artifact" or "tool". Next, we want to examine problems which could arise if AI was categorized like a non-artificial "moral agent". Whether we count a person morally accountable or not affects the way we expect to be treated by then and how we treat them in return; for the judicial institution to function, we must assess legal accountability, which in turn is partially grounded on moral accountability. We seem to have an inclination to regard AI as a moral agent [7, 177], although people do not consider them as appropriate agents for moral decisions (see [115, 139]). Usually, when encountering intelligent and goal-directed behavior, the human mind attributes the target with core features of agency, that is, the presence of a self and consciousness [181]. In other words, we assume a conscious being to be behind intelligent action; or, at the very least, we assume intelligent action to be caused by some kind of intentional (volitional and feeling) goal-oriented self. Such inferences are a useful cognitive strategy in an environment populated mostly by other humans, as it has been the case for *Homo sapiens* for thousands of years. When we encounter goal-oriented behavior, our folk psychology systems become activated and we start making inferences about the agents' desires, thoughts, feelings, and beliefs.

However, folk psychology becomes a hindrance when we apply it to AI. AI systems might behave intelligently in the sense that they are capable of goal-oriented action in some specific environments. Despite this, they are generally thought to lack consciousness and an experiential sense of themselves or the world: nothing "feels" or "appears" as something to them [79]. We do not want to take part in the wide philosophical discussion about consciousness and the possibility of artificial consciousness too deeply (see [16]). We are content to remark that, at least the current systems have no selfhood or agency similar to non-artificial agency, let alone anything resembling consciousness as we know it. However, these absences in no way eliminate the possibility of intelligence. A system can be intelligent without being conscious, and similarly, a being can be conscious without being particularly intelligent (e.g., a human baby). In these respects, AI resembles animals. Even if an animal would be regarded incapable of consciousness, it could still be regarded intelligent, according to the cognitive science definition. Cognitive science has not set a limit to what kind of entity can be intelligent and even plant intelligence and learning has been studied [58]. The key message is that intelligence is not just a property of the human brain, and that intelligent action is not required to be conscious.

To summarize superficially, what makes the situation complicated is that there is no philosophical or scientific consensus on the nature, origin, or function of consciousness. The debate has been going on since the 1970s [16]. Some philosophers, such as Daniel Dennett [41], have proposed that in principle, there is no difference between human and artificial consciousness. If consciousness is information processing in the brain, an artificial being can be capable of it as well. It is also often generally thought that selfhood and consciousness are required for full moral agency [142, 177], but according to this view, there is no selfhood separate from the cognitive system and the brain. However, the line of thought represented by Dennett has its critics. For example, mathematician Roger Penrose [144], neuroscientist Giulio Tononi [170], and philosopher David Pearce [143] have proposed that some of the computational properties of consciousness might not be implementable on a silicon-based microchip or on a Turing machine architecture. If they are right, technologies based on classical computation and microchip architecture could not become conscious, not to mention have moral agency or a self. This applies to all currently designed AIs.

There is more to the question whether AI has true moral agency than simply whether it can be conscious. Moral agency is better viewed as a spectrum. Even if AIs were not conscious, but still behaved similarly enough to humans, they could perhaps be counted as moral agents of some kind [142, 177]. This functional equivalence could then form the basis for at least some kind of moral responsibility and AIs might thus be placed on the moral agency spectrum. For instance, there has been discussion on the foundation of moral responsibility [71]. However, philosophers have been very skeptical about placing AI systems high or even to the middle on the moral agency spectrum [9, 28].

Robots and AIs can clearly be intelligent, that is, they can perform tasks and act in goal-oriented ways in an at least partially predictable environment. At the same time, they are not capable of extensive moral agency [177], so they can hardly be held morally or legally accountable any time in the near future. However, it does not follow that they would be morally neutral or irrelevant, since their (intelligent) actions can have an impact on people's well-being. Even a piece of simple technology, such as a hammer or a baseball bat, is not morally neutral. Instead, it reflects the values and objectives of its designers and its chain of production [55, 150]. This is even more evident with AI, since it blends into our social and moral lives, and is more interactive, autonomous and social than hammers or baseball bats.

7 The doctrine of double effect and the problems of folk consequentialism

We now turn away from challenges related to categorization and move to more general limitations of human moral cognition. Various simplifications and biases steer the conversation regarding AI ethics towards over-generalization and dead-ends [104, 105, 116, 127, 177]. As philosophers have noted during the millennia-long debate, it is challenging to establish uniform and consistent stances on ethical problems and the basic issues of ethics (see [53]). Many debates in philosophical ethics take place between incommensurable normative theories, such as consequentialism, deontological ethics and virtue ethics. Understanding these debates often requires delving deep into the conversation or even getting an education in philosophy. Our everyday moral thinking is often simpler and grounded on our social and moral emotions [68, 69]. Folk ethics typically simplifies complex moral problems and hides their nuances. This human tendency results in many difficulties in the context of democratizing AI ethics.

One of the central theories in normative ethics claims that moral acceptability of actions depends only on their consequences [40]. Consequentialism comes in several flavours, which differ based on how they evaluate the costs and benefits associated with the consequences of different actions [160]. Consequentialist arguments have been especially strongly represented in the discourse about new technologies and their introduction into our society [62, 63, 153]. For example, many arguments for autonomous traffic are based on the claim that it ultimately saves human lives. It is, therefore, justified, the argument goes, to test the autonomous traffic technology in the public space and risk a few deaths.

This kind of one-sided folk consequentialist reasoning is riddled with apparent difficulties. Automatized traffic with self-driving cars has to be planned and designed in advance of implementation, which means that the designers are forced to make choices on who is expendable in potential accidents [7, 19]. This reveals one of the best-known counterarguments against consequentialism: the estimation of utility is extremely difficult (who receives the benefit, at what price, with what externalities, etc.). We can make an additional remark that consequentialism clashes with some of our other moral intuitions, if the maximization of utility is applied straightforwardly towards the moral evaluation of actions. Our moral intuitions estimate the moral value of actions themselves (e.g., stealing is wrong) not only on their beneficial results, but (among other considerations) also based on the moral norms and values involved. For example, according to some forms of consequentialism, if someone's internal organs could save ten other people, the harvesting and reuse of this person's internal organs would be regarded as morally right regardless of the person's will (for a discussion on this classical argument, see [56], pp 19–32). Studying consequentialist thinking empirically has turned out to be difficult due to challenges in operationalizing such views [98].

Even if we manage to avoid the challenges of a priori utility estimation (and definition of utility), other problems can arise with respect to AI ethics. Problems arise, for example, when consequentialism based on short-term benefits is indiscriminately applied (without considering other values, such as human dignity, responsibilities, and rights). Consider again the case of automatized traffic. Autonomous cars could be designed to protect its passenger at the expense of a pedestrian in the event of an accident (the alternative would likely not sell). In this case, the fact that someone has the money to buy this particular vehicle becomes the factor deciding between life or death, raising the passenger's life to a higher value than that of the pedestrian. This banal example shows how equal human rights can be overridden by vulgar consequentialism, while technology escalates inequalities across the board [83]. For these reasons, it is crucial that the discussion about AI is not restricted to a narrow ethical perspective.

The successful application of consequentialism within AI ethics is further hindered by cognitive biases recognized in moral psychology. One of these is the doctrine of double effect (DDE) [121, 165].[9] People consider killing someone to save another person more morally acceptable when it is done by pushing a button or pulling the rope of a guillotine (that is, indirectly or instrumentally), compared to directly strangling someone with their "own hands". The consequences of the action are identical (i.e., someone dies). We are susceptible to the doctrine of double effect when we consider self-driving cars in the folk ethics context. Double effect research findings suggest that we do not realize the gravity of the ethical issue when a device or a gadget (e.g., a car) is part of the chain of causation, because it dilutes our perceptions regarding personal agency as a part of the events in question [100, 126]. In a sense, the general human tendency towards the DDE is a form of perceptual bias in the area of moral perception.

When a person drives a car and ends up in a fatal accident, figuring out who is responsible is a comparatively easy process. If the driver caused the accident, they can be punished within some boundary conditions. In such a case,

[9] Uwe Steinhoff comments on the "doctrine of double effect" (DDE) in his recent analysis [165]. He argues that the phenomenon is a bias by saying that "The methodology [i.e., conceptual analysis/argument] used by defenders of the DDE or related principles is driven by bias and it is deeply flawed." We refrain from further commenting on this heated debate here.

third parties find the punishment natural and consistent with the general moral views prevalent in our society. On the other hand, if the car is autonomous, the matter of allocating responsibility becomes more complicated. It seems that responsibility would be distributed over a network of actors, whose status and contribution to the event is somewhat unclear. This could result in a "responsibility gap". Does the responsibility lay with the car itself, the programmer of the car's driving algorithms, the manufacturer of the car, or society more generally? Who should be punished in such a situation? Probably the upper management of the company that produced the car would not be held morally responsible. Quite likely the company that developed the car will be held at most only partially responsible and let off the hook with a fine for negligence. This, in turn, might lead to valuing human lives increasingly in economic terms, the less privileged facing a bigger risk of losing their lives, and the family and friends of the deceased having a lasting sense of injustice and a lack of closure [48].

With the spread of autonomous vehicles, publicly funded roads might become a kind of test laboratory for private companies, on which the consequences of accidents can simply be bought off. In the worst case, vulgar consequentialism might lead to a situation in which, when we enter a public space, we involuntarily become product testers for new technology—not unlike crowd sourced crash test dummies. New technology might be developed in the name of consequentialism and escorted with the promise of increasing safety, while actually motivated by profit or control (the latter most obvious in the case of China). This would in turn lead to rationalizations on why compromising human rights is justified. No university ethical review board would approve such a large-scale pseudoscientific experiment, but the general atmosphere appears to be more permissive with respect to corporate-based R&D.

Recent research suggests that the above worries are not ungrounded. Studies have shown that people prefer self-driving cars to function according to simple consequentialist principles, that is in a way which maximizes the number of human lives saved, even if it meant sacrificing the passengers [7, 19]. This appears to be an encouraging altruistic result. However, some of the same studies further show that people's preferences are egoistic [19]: they prefer others to use cars following consequentialist principles, while preferring not being passengers in such cars themselves. In other words, people endorse consequentialism until they are required to sacrifice themselves for the benefit of others, in which case even vulgar folk consequentialism becomes too high a bar.

To manage the risks associated with vulgar folk consequentialism, egoism and the doctrine of double effect, we need a wide consensus and adherence to human dignity, democratically determined agreements, rights and responsibilities. There is a general need to regulate the development of technology collectively to ensure that it serves the common good and supports the rule of law. Otherwise, the risk grows in the future for that technology to become an uncontrolled profit maximizing power that forcefully instrumentalizes humans. Fortunately, the Western tradition of philosophy, theology, and political thought provides us with abundant resources to counter the effects of vulgar consequentialism. Governance by the rule of law, where an individual has absolute value that cannot be transformed into money or utility, is a key institution for the development of ethical technology.

8 Values beyond safety

Safety is a major theme of AI ethics. However, aiming towards safety can also predispose us to biases stemming from our everyday thinking. Safety is a moral principle among others and its prioritization might lead to deeply problematic outcomes. It seems that when AI is marketed with themes of increased safety and fear, the perceived threat often stems from other people—we will elaborate on this below.

The 2002 film Minority Report portrays an information-processing system which can predict possible murders and other crimes. The technology is used to prevent all crimes in advance and the world is perfectly safe, superficially. A world, where no wrong is ever made appears highly desirable and good. The film explores the problem of false positives: many people are preemptively jailed even when they never actually committed a single crime. In principle, each of us is a potential criminal. Likewise, no one would commit crimes if, as a precaution, everyone was put in permanent solitary confinement. However, we have no way of knowing if a crime would have necessarily been committed, until it was actually committed. We can only estimate the probabilities of some event taking place. There are many activities which superficially resemble the preparation of a crime, while they are actually absolutely harmless (e.g., the growing of vegetables under a heat lamp[10]).

In light of research in personality and moral psychology, people can be characterized along a continuum (broadly speaking) as: open-minded and adventurous, or cautious and conscientious [3, 8, 21, 31, 43, 49, 57, 80, 113, 120, 133, 135, 155, 187]. Open-minded and adventurous people form a minority in our society, but they produce the majority of our new ideas, inventions and innovations. Cautious

[10] https://www.dailymail.co.uk/news/article-2301308/Ex-CIA-agents-sue-Kansas-police-raided-suburban-home-drugs--bought-special-equipment-grow-vegetables-indoors.html.

and conscientious people, on the other hand, are responsible for the functioning and maintenance of society and institutions. These personality clusters are also aligned with different views on everyday moral decisions [34]. Open-minded and adventurous people tend to view moral acts and societies primarily through the lenses of (i) fairness, and (ii) whether such actions or political decisions cause harm to someone. Cautious and conscientious people, in turn, tend to view some other considerations as also relevant for moral judgement: they consider moral actions and political decisions through the lenses of respecting (iii) the local societal norms, (iv) the local societal authorities, or (v) values locally regarded as sacred (whatever such values happen to be; [64, 70]). Cautious and conscientious people are usually more sensitive towards feelings of disgust and fear than open-minded and adventurous people [172], and in turn people with lower sexual disgust sensitivity rely more on consequentialist thinking [99, 102].

Probably due to this predominance of the cautious personality style, many AI systems and emerging technologies are marketed with overtones of fear and safety [15, 26, 59, 76, 129, 186]. These technologies include face recognition algorithms for automated surveillance and profiling [140]. Profiling data in itself was deemed unfair evidence in court: the defendant would be seen responsible for the crimes of others and not their own, but such technology does make certain people more likely law enforcement targets.[11] Another such technology, already implemented in the US, predicts in which neighborhoods a crime is most likely to occur.[12] Police forces are channeled into the neighborhood, leading to self-fulfilling prophecies not unlike in Minority Report: the police strive to apprehend anyone who might have cannabis in their pocket, because such crimes are found to be more frequent in that area, because the police successfully apprehends people with cannabis in their pockets.

Without education, people will have a hard time understanding AI ethics and the relevant sociological issues. Populations unfamiliar with the properties of AI systems can easily be persuaded to, "for their own safety", implement technologies which, actually, might erode the very prerequisites of constitutional democracy [45].

9 Our crimes are not really crimes, but theirs are

The evolution of our moral and social mind took place in the context of relatively small and competitively antagonistic social groups [22, 162]. This is considered to be at least a part of the explanation for the unique and throughout social nature of the human mind [169]. One of the mental tendencies stemming from this social origin is the human tendency to split people into in-groups and out-groups [52]. Individuals categorized in the in-group are perceived as more valuable than those categorized as out-group members. We could simply call this the *us and them bias*; and it too distorts the discussions taking place around AI policy. Algorithmic profiling is one application, where this bias becomes readily apparent. AI makes possible novel ways for detecting crime such as through information tracking. Certain kinds of crimes are associated with certain kinds of social groups. People are likely to use stronger measures to deter and punish crimes associated with an out-group, and conversely be more cautious and lenient when considering measures to track crimes perceived to be more common within their in-group [112, 117].

The profiling system for preventing drug related crimes, introduced above, is less capable of preventing financial crime, which, on the whole, leads to significantly higher costs and damage to society. Financial crime is not intuitively perceived as a societal security risk, even when it detracts from the financial foundation of society, and in the long term causes higher rates of alcoholism, family violence, and suicides [45]. There is at least one recent example of applying AI algorithms to counter tax evasion and other financial crimes within the EU [173, 174]. When the algorithm was ready, it was run through the tax records of the state. The goal was to extract thousands of names for observation and investigation, but the Euro-group and troika[13] obstructed the execution of this program [172, 173]. It is unlikely that similar EU governance level objections would be raised if big data was used to counter crime, such as local drug use. To raise a potential example, in theory it would be possible to create a database which combines individuals' musical preferences to their travel data and the chemical analyses of the wastewater in their residential area to predict drug use.

One of the common stereotypes of a drug user is that of a poor person who is in risk of becoming a pariah [61, 185]. However, international research shows that most people who have used drugs are never caught, and their drug use has no identifiable negative effects on the rest of society [132]. Illegal drugs are used by the well-off members of

[11] https://blog.arizonacriminaldefenselawyer.com/when-drug-courier-profiling-evidence-is-used-against-you-at-trial-for-the-purpose-of-proving-guilt-it-deprives-you-of-an-essential-right/.

[12] https://www.technologyreview.com/s/612957/predictive-policing-algorithms-ai-crime-dirty-data/.

[13] https://en.wikipedia.org/wiki/European_troika.

society, such as physicians, lawyers, clinical psychologists, university professors, IT entrepreneurs, physicists, primary school teachers, and social workers [132]. However, different demographics tend to use different drugs or commit specific drug related crimes in particular [137]. If it suited the intentions of the government, the lives of thousands of citizens associated with drug related crime could be made more difficult through the analysis of the above-mentioned sensitive information.

These attempts to apply AI to crime prevention clearly show the *us and them bias*: the in-group and its proclivities are held in higher regard and are more worthy than those of the out-group (see [175]). This bias becomes more pronounced when AI is used to fight crime committed by people who do not belong to the same in-group as the policymakers, most obviously the poor and the immigrants.[14] Most of us tend to assess the justifications of moral actions based on whether such actions adhere to the values and norms of our own native culture, and we might be ready to support the implementation of profiling algorithms in the fight against crime. People consistently find it difficult to imagine that they themselves are a part of some out-group and would thus be the target of surveillance technology. Implementing new "security providing" technologies is seen as a great idea—until the technology is turned against oneself.

A surveillance system capable of ethnic or other profiling can easily be calibrated to target new groups. Does it lead to a better world if human action can be surveyed and controlled so precisely? It might appear obvious that in a democratic society more efficient law enforcement is beneficial, but the downside to this would be that beneficial social functions which interact with illegal activities would have to adapt to new circumstances. As law enforcement becomes more efficient, so too the laws need to become more nuanced. The moral and ethical advances of society are often marked by innovations occurring in the "moral gray area" [27]. Two drug-assisted treatments have recently been granted the Breakthrough Therapy designation by the US Food and Drug Administration (FDA). FDA assessed two Schedule I substances to have significant potential in treating several severe mental health disorders.[15] These treatments would unlikely have been developed without the dedicated advocacy of activists belonging to socially marginalized groups [187]. To conclude this section, while we examined profiling algorithms solely from an us-them perspective, this is not to say that such technologies are not fraught with ethical problems related to privacy and other human rights.

10 Egocentric teleology bias

Here we introduce a concept that has only recently surfaced in the cognitive bias literature [146]. In essence, egocentric teleology is an amalgamation of several folk theories that neatly encapsulates several themes under the same umbrella. It has two parts, which we will cover in order.

Egocentricity refers to our tendency to view the world from our own perspectives. Everyone has an intuitive sense of themselves being quite a complex and in some way unique agent; furthermore, we tend to feel that we personally are somehow special and exceptional when compared to other humans [146]. In feeling special, we also as if the surrounding world was somehow there personally *for* us, and adheres to our personal wishes and needs.

Teleology (from Greek *telos*, "end", and *logos*, "reason") is traditionally described as the explanation by reference to purpose, end, goal or function. Humans often have teleological reflections on the behaviour of in nature, and also see themselves as pursuing ends and goals. Some experimental studies show that children seem to project function and design to the natural world [46, 90]. In these studies, preschool-age children attribute, for example, goal-directed actions to tigers, icebergs and rocks. Even PhD-level physical scientists and their work are not immune to intuitive and teleological explanations of natural phenomena [92, 93].

This egocentric view is also reflected in the teleological perspective according to which objects, whether human-made or natural, are designed for their appropriate human goals and through this design the objects play a valuable function, which in turn makes those objects *good*. This view is suggested to be a generalization from the intuition that some objects or phenomena appear to serve a particular function and are too complex to exist by pure chance [146]. Without scientific knowledge, the human eye seems to be finely suited for seeing and too complex to have come through any process but design. By extrapolation, chairs are designed for sitting, as are large flat rocks; and spears for hunting, as are small sharp rocks. Everything in the world appears to have a purpose. Even events and circumstances are what they are because they happened for our benefit and lead up to our well being, since they are designed for us by some higher power [146]. Similarly, more complex technologies appear to have come about, to be designed, for their own designated purposes—"otherwise how could something so unlikely have come about". When it comes to technologies

[14] https://www.theguardian.com/technology/2019/oct/16/digital-welfare-state-big-tech-allowed-to-target-and-surveil-the-poor-unwarns; https://www.wired.com/story/opinion-ai-for-good-is-often-bad/; https://www.cigionline.org/articles/using-ai-immigration-decisions-could-jeopardize-human-rights.

[15] https://www.medscape.com/viewarticle/921789; https://www.healthline.com/health-news/fda-looking-at-magic-mushroom-ingredient-to-treat-depression; https://www.nytimes.com/2018/05/01/us/ecstasy-molly-ptsd-mdma.html.

used for apparently benign purposes, the implicit assumption is that due to their design and purpose, they cannot be the tools ultimately causing adversity.

The egocentric teleology bias has had a part in shaping human history: for example, other animal species have been viewed to exist for the sole purpose of human exploitation, either as food or auxiliary labour [146]. The egocentricity bias can be a part of a complete worldview, where the whole of reality (and especially other animals and plants) is perceived to exist for humans and to be at human disposal. Closer to home, the egocentric bias can be observed in the general human tendency to regard themselves as goal-oriented, complex, and good. By proxy of "knowing" their own goodness, all groups in which a person belongs are thus also, in some sense, good. A good person cannot belong to a bad group, "and even if others can end up in bad company, I surely cannot". This egocentric tendency has time and time again led to a phenomenon called dehumanization [74, 75, 180] slaves are not human, and Tutsis are cockroaches (sic; [152]).

We tend to see all complicated and superficially goal-oriented technologies as ethically less problematic than they might actually be, in part due to the teleology bias. Smartphones, cars, televisions, nuclear power stations, profiling algorithms, and, for example, care robots might all be intuitively perceived as morally neutral, or even as progressive and morally good. We usually do not stop to reflect on cars and smartphones in themselves as morally relevant objects even if their use and influence on the world might be analyzed from a moral standpoint. In actuality, technological tools cannot be disentangled from moral perceptions: they are not produced in a moral vacuum. As many geologists have pointed out, we have now entered the Anthropocene [138], a new geological era in which Earth is actively reshaped according to short-term human needs. The whole surface of the world is being changed due to technological developments and large portions of the changes are deleterious for the long-term survival of humans on this planet. This means that the technological developments are hostile to life and thus morally questionable, if—given the results—not outright evil.

The egocentric teleology bias is one of many factors which blinds us morally to the challenges of novel technologies. We recognize robots as somewhat autonomous goal-oriented agents, while at the same time, they appear to function as if by magic. Even with their fairly primitive neural networks, robots are already too complicated for us to understand the way they function [127]. We have an intuitive sense that complex goal-oriented technologies, including robots, by the virtue of having been developed by us, are morally good, or at least neutral, when it is just as likely that they are not. We suggest that this moral optimism is in part due how we perceive ourselves: goal-oriented, complex, and good. We perceive AI technologies as extensions of and sharing our traits and are hesitant to doubt their moral value. Deliberate and careful philosophical consideration and inclusive ethical discussion is crucial to the proper maturation of complex novel technologies.

However, the egocentric teleological bias is not the only cause leading us to view technological systems as good. They are perceived as morally good also because the theoretical models on which such systems rely are perceived as "rational" and as something that accurately correspond to the structure of reality, and are thus further perceived as a viable solution to the threat posed by "The Others" (i.e., out-group members, see above). For example, some game theoretical models have been used in designing the automatization of some functions of social institutions [45]. It is crucial to realize that such models are limited to simulating human social behavior without taking into account all relevant non-social factors.

11 Biases of wish fulfillment in risk estimation

We have examined many predicted and actual cases in which AI threatens our values. Regardless of these examples, AI, as a technology, promises unprecedented wish fulfillment. We now briefly examine some of our biases in examining the moral worth of a technology of such a magnitude as AI. The argument developed here is not specific to AI but also similar vaguely defined socially influential scenarios. The scope of problems that could be solved or dreams realized with the right kind of AI technology is vast (unprecedented even), not least due to the potential of AI to act as a "meta-technology" accelerating the development of other revolutionary technologies [2]. In the best case, AI could transform our society in positive ways surpassing anything we could achieve otherwise [89]. Analogous thinking has been applied to AI as to the Christian promise of Heaven and eternal delight surpassing anything conceivable during one's mortal existence, and in this sense, AI most resembles the category of *god* (already with its first Church[16]). As with Pascal's Wager, from a strict utility maximization perspective, pursuing the scenario with infinite expected utility is the rational choice regardless of its probability [20, 82]. Similarly, any action which promotes the development of utopian AI would appear rational regardless of the probability of achieving the desired outcome, whatever the sacrifices.

[16] https://www.wired.com/story/anthony-levandowski-artificial-intelligence-religion/.

Such thinking might rightly leave professional and folk economists unconvinced [130]. However, our intuitive cognitions can be hypothesized to react with credulity to near infinite rewards and highly desirable scenarios even when their estimated likelihood is extremely slim and they border on the fantastic (i.e., people wish that magic was true with respect to AI technology). Attitudes towards objects, people, or scenarios develop in part through association rather than analytical reasoning (see [18]. Not only the truth matters but also the story. What desires could AI fulfill? Personalized and engaging education or technical training in any skill[17]; crime-free society; dream fantasy worlds of adventure, excitement, sex, and fame [73], a post-scarcity economy [39], world peace[18]; near unlimited energy [89]. Encountering these and countless other plausible and less plausible ideas can add up to a general sense that AI is awesome.

Would not all the examples of dystopian consequences of AI balance out the utopian examples? Not necessarily. We tend to view technologies as either all-good or all-bad. If we encounter information which suggests higher than expected benefits of a technology, we consequently downplay the risks, and vice versa if we gain information relating to the risks. This is, of course, contrary to how higher benefits and risks are generally actually coupled: the larger the benefit, so also larger the risk [51, 161]. Even scientifically minded people remain susceptible to these biases after receiving further content or context knowledge [85]. Additionally, research suggests that we find our preferred scenarios more likely and update our beliefs asymmetrically by giving more weight to evidence in line with our wishes [84]. It is plausible that some people are cautious towards many or most AI applications, but due to being highly optimistic about a particular application which appears more salient to them personally (e.g., from the perspective of safety), are perhaps overoptimistic and insufficiently cautious about AI development in general. Science fiction hobbyism predicts more positive attitudes towards robots regardless of whether the sci-fi portrays robots in a positive (Star Trek) or negative light (Terminator) [95, 96, 100, 102]. The mere presence of extreme wish fulfillment scenarios may bias us to approach a scenario less cautiously.

The promise of high rewards increases significant risk taking. Some of the most admired people dropped out of college and college students dream of founding successful start-ups rather than completing their studies. However, the tendency to overestimate the chances of success based on the saliency of success stories is a classic case of survivorship bias.[19] This is exacerbated by our tendency to make more risky choices when we feel positive [33] (see also [67]). It might be socially beneficial that many of us attempt to achieve something in which a tiny fraction succeeds, but no such similar strategy serves us if we think how to guide society itself.

Risk taking motivated by high rewards can also be an act of omission. We might be tempted to withhold from acting in the belief that exponentially developing technology will solve our problem in a short while. The idea of rapid medical advances might decrease the motivation to develop an exercise habit; hopes of an automated society could decrease the felt need for learning marketable skills and building a reputation. Similarly, on a larger scale, emerging technologies are relied upon to solve our increasing energy demands [182] or the risks of climate change [77]. While they are surely a part of the solution, their presence on the horizon, however unlikely, can lead us to misjudge our priorities. The challenge is plotting a course to daring beneficial AI applications without relying on miracles.

12 Conclusion

Whether we want it or not, humanity has evidently entered a new era in which we face novel moral challenges. Such challenges are unprecedented in both their deep and surface structures [72]. For the first time in human history, we find ourselves in an environment in which designed lifeless and unconscious matter makes decisions which deeply affect human wellbeing. Our cognitive limitations and biases might lead the development of such technology into a direction, where its consequences are not desirable, not only in a totalitarian but also a free democratic society.

Through the shifts we are undergoing, technology increasingly functions as the mirror of our moral cognition and political systems. We can consciously guide the development of technology into a direction, where it supports our democratic systems and promotes the principles of equality and justice; or we can allow the end result to be based on market forces and innate human biases. Our moral cognition is shaped by both our emotions and by our primitive herd instincts. If we want to avoid a world shaped by the instinctual animalistic needs of our nature, we need to take into consideration the motivational forces and possible stone age biases that could inadvertently guide the development of AI, or prevent wider discussion about it.

[17] https://www.nytimes.com/2019/12/18/education/artificial-intelligence-tutors-teachers.html; https://www.lesswrong.com/posts/vbWBJGWyWyKyoxLBe/darpa-digital-tutor-four-months-to-total-technical-expertise.

[18] https://www.aljazeera.com/features/2017/5/30/could-artificial-intelligence-lead-to-world-peace.

[19] https://www.forbes.com/sites/carminegallo/2012/12/06/high-tech-dropouts-misinterpret-steve-jobs-advice/.

In this text, we have described (depending on how they are counted) around 14 different features of human cognition which in part explain why discussing AI technologies and their risk factors is so difficult. The set of intertwined problems described above stems from the fact that humans evolved in an environment, where there were no AI agents, and thus we do not have innate concepts for probability and conditionality. Humans are social animals that use their intuitive and automatic cognitions to understand the social world specifically [163]. To function in their social environment, we tend to egocentrically project our humanness and conscious agency into our environment and especially on AIs. AI technology is not something that we grasp intuitively through our categorical cognition. We hope that our text gives food for thought to those who are working in relevant fields and encourages the discussion on the social risks of AI, especially when addressing the general public, and reveals potential new directions. As humans, we are facing the choice of having enlightened and cautious discussions by which we prepare to face the risks of AI consciously and guide humanity to a more democratic and egalitarian society, such as in Star Trek, or not having such civilized discussions and end up in a situation more resembling the dystopias of Blade Runner or The Matrix.

Acknowledgements The authors would like to thank Jane and Aatos Erkko Foundation (grant number 170112) and the Academy of Finland (Grant number 323207) for their funding.

Funding Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Agrawal, S., Williams, M.A.: Robot authority and human obedience: A study of human behaviour using a robot security guard. In: Proceedings of the companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 57–58 (2017)
2. Agrawal, A., McHale, J., Oettl, A.: Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth (No. w24541). National Bureau of Economic Research (2018)
3. Aluja, A., García, O., Garcia, L.F.: Relationships among extraversion, openness to experience, and sensation seeking. Personality Individ. Differ. **35**(3), 671–680 (2003)
4. Amiot, C.E., Bastian, B.: Toward a psychology of human-animal relations. Psychol. Bull. **141**(1), 6–47 (2015)
5. Atran, S.: Modular and cultural factors in biological understanding: an experimental approach to the cognitive basis of science. In: Carruthers, P., Stich, S., Siegal, M. (eds.) The cognitive basis of science, 41–72. Cambridge University Press (2002)
6. Atran, S., Medin, D., Ross, N.: Evolution and devolution of knowledge: a tale of two biologies. J. R. Anthropol. Inst. **10**(2), 395–420 (2004)
7. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)
8. Baer, M., Oldham, G.R.: The curvilinear relation between experienced creative time pressure and creativity: moderating effects of openness to experience and support for creativity. J. Appl. Psychol. **91**(4), 963–970 (2006)
9. Banks, J.: A perceived moral agency scale: Development and validation of a metric for humans and social machines. Comput. Hum. Behav. **90**, 363–371 (2019)
10. Baran, B.E., Rogelberg, S.G., Clausen, T.: Routinized killing of animals: Going beyond dirty work and prestige to understand the well-being of slaughterhouse workers. Organization **23**(3), 351–369 (2016)
11. Barratt, J.: Our Final Invention. Macmillan (2013)
12. Barrett, J.L.: Exploring the natural foundations of religion. Trends Cogn. Sci. **4**(1), 29–34 (2000)
13. Barrett, J.L.: Born Believers: The Science of Children's Religious Belief. The Free Press, New York (2012)
14. Bigman, Y.E., Gray, K.: People are averse to machines making moral decisions. Cognition **181**, 21–34 (2018)
15. Bird, S, Tapp, A.: Fear and fire: ethical social marketing strategies for home fire safety for older people. Retrieved February 2, 2021, from https://uwerepository.worktribe.com/output/963462. (2011)
16. Blackmore, S., Troscianko, E.: Consciousness: An Introduction, 3rd edn. Routledge, London (2018)
17. Boden, M.: Mind as Machine: A History of Cognitive Science. Oxford University Press (2008)
18. Bohner, G., Dickel, N.: Attitudes and attitude change. Annu. Rev. Psychol. **62**, 391–417 (2011)
19. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
20. Bostrom, N.: Pascal's mugging. Analysis **69**(3), 443–445 (2009)
21. Bouso, J.C., Palhano-Fontes, F., Rodríguez-Fornells, A., Ribeiro, S., Sanches, R., Crippa, J.A., Hallak, J., Barros de Araujo, D., Riba, J.: Long-term use of psychedelic drugs is associated with differences in brain structure and personality in humans. Eur. Neuropsychopharmacol. **25**(4), 483–492 (2015)
22. Boyd, R., Richerson, P.J.: The Origin and Evolution of Cultures. Oxford University Press (2005)
23. Boyer, P.: Religion Explained: The Evolutionary Origins of Religious Thought. Basic Books, New York (2001)
24. Boyer, P., Barrett, C.: Evolved intuitive ontology: integrating neural, behavioral and developmental aspects of domain-specificity. In: Buss, D.M. (ed.) Handbook of Evolutionary Psychology. Wiley (2005)

25. Breazeal, C., Gray, J., Hoffman, G., Berlin, M.: Social robots: beyond tools to partners. In: *RO*-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), 551–556 (2004)
26. Brennan, L., Binney, W.: Fear, Guilt, and Shame Appeals in Social Marketing. J. Bus. Res. **63**(2), 140–146 (2010)
27. Brownlee K.: Civil disobedience. In: Zalta E.N. (ed.) The Stanford Encyclopedia of Philosophy. Retrieved February 2, 2021, from https://plato.stanford.edu/archives/fall2017/entries/civil-disobedience/ (2017)
28. Brożek, B., Janik, B.: Can artificial intelligences be moral agents? New Ideas Psychol. **54**, 101–106 (2019)
29. Cappuccio, M.L., Peeters, A., McDonald, W.: Sympathy for Dolores: moral consideration for robots based on virtue and recognition. Philo. Technol. **33**(1), 9–31 (2020)
30. Carlson, Z., Lemmon, L., Higgins, M.C., Frank, D., Salek Shahrezaie, R., Feil-Seifer, D.: Perceived mistreatment and emotional capability following aggressive treatment of robots and computers. Int. J. Soc. Robot. **11**, 727–739 (2019)
31. Carney, D.R., Jost, J.T., Gosling, S.D., Potter, J.: The secret lives of liberals and conservatives: personality profiles, interaction styles, and the things they leave behind. Polit. Psychol. **29**(6), 807–840 (2008)
32. Castelvecchi, D.: Can we open the black box of AI? Nature News **538**(7623), 20 (2016)
33. Cheung, E., Mikels, J.A.: I'm feeling lucky: The relationship between affect and risk-seeking in the framing effect. Emotion **11**(4), 852 (2011)
34. Clark, C.B., Swails, J., Pontinen, H.M., Bowerman, S., Kriz, K.A., Hendricks, P.S.: A behavioral economic assessment of individualizing versus binding moral foundations. Pers. Individ. Differ. **112**, 49–54 (2017)
35. Coeckelbergh, M.: Humans, animals, and robots: a phenomenological approach to human-robot relations. Int. J. Soc. Robot. **3**(2), 197–204 (2011)
36. Coghlan, S., Vetere, F., Waycott, J., Barbosa, N.B.: Could Social Robots Make Us Kinder or Crueller to Humans and Animals? Int. J. Soc. Robot. **11**(5), 741–751 (2019)
37. Cormier, D., Newman, G., Nakane, M., Young, J.E., Durocher, S.: Would you do as a robot commands? An obedience study for human-robot interaction. In: International Conference on Human-Agent Interaction (2013)
38. Cosmides, L., Barrett, C., Tooby, J.: Adaptive specializations, social exchange, and the evolution of human intelligence. Proc. Natl. Acad. Sci. **107**(Supplement_2), 9007–9014 (2010)
39. Danaher, J.: Automation and Utopia: Human Flourishing in a World Without Work. Harvard University Press (2019)
40. Darwall, S.: Consequentialism. Blackwell, Oxford (2003)
41. Dennett, D.: Consciousness Explained. Penguin (1992)
42. Dennett, D.: Freedom Evolves. Penguin (2003)
43. Ebstein, R.P., Monakhov, M.V., Lu, Y., Jiang, Y., Lai, P.S., Chew, S.H.: Association between the dopamine D4 receptor gene exon iii variable number of tandem repeats and political attitudes in female Han Chinese. Proc. R. Soc. B Biol. Sci. **282**, 20151360 (1813)
44. Erickson, T.D., Pickover, C.A., Vukovic, M.: U.S. Patent No. 10,683,088. Washington, DC: U.S. Patent and Trademark Office (2020)
45. Eubanks, V.: Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press (2017)
46. Evans, E.M.: Cognitive and contextual factors in the emergence of diverse belief systems: creation versus evolution. Cogn. Psychol. **42**(3), 217–266 (2001)
47. Evans, J.S.B.: In two minds: dual-process accounts of reasoning. Trends Cogn. Sci. **7**(10), 454–459 (2003)
48. Federico, C.M., Ekstrom, P., Tagar, M.R., Williams, A.L.: Epistemic Motivation and the structure of moral intuition: dispositional need for closure as a predictor of individualizing and binding morality. Eur. J. Pers. **30**(3), 227–239 (2016)
49. Feist, G.J., Brady, T.R.: Openness to experience, non-conformity, and the preference for abstract art. Empir. Stud. Arts **22**(1), 77–89 (2004)
50. Ferreira, C.M., Serpa, S.: Rationalization and bureaucracy: Ideal-type bureaucracy by Max Weber. Hum. Soc. Sci. Rev. **7**(2), 187–195 (2019)
51. Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M.: The affect heuristic in judgments of risks and benefits. J. Behav. Decis. Mak. **13**(1), 1–17 (2000)
52. Fiske, S.T., Taylor, S.E.: Social Cognition: From Brains to Culture. Sage (2013)
53. Frances, B., Matheson, J.: Disagreement. In: Zalta E.N. (ed.) The Stanford Encyclopedia of Philosophy. (Winter 2019 Edition). Retrieved February 2, 2021, from https://plato.stanford.edu/archives/win2019/entries/disagreement/ (2019)
54. Frischmann, B., Selinger, E.: Re-engineering humanity. Cambridge University Press (2018)
55. Friedman, B., Kahn, P.H., Jr.: Human values, ethics, and design. In: Jacko, J. (ed.) Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, pp. 1209–1233. CRC Press (2002)
56. Foot, P.: Virtues & Vices. Oxford University Press, Oxford (2003)
57. Furnham, A., Crump, J., Batey, M., Chamorro-Premuzic, T.: Personality and ability predictors of the "Consequences" test of divergent thinking in a large non-student sample. Pers. Individ. Differ. **46**(4), 536–540 (2009)
58. Gagliano, M.: In a green frame of mind: perspectives on the behavioural ecology and cognitive nature of plants. AoB Plants **7**, plu075 (2015)
59. Gagnon, M., Jacob, J.D., Holmes, D.: Governing through (in) security: a critical analysis of a fear-based public health campaign. Critical Public Health **20**(2), 245–256 (2010)
60. Geerdts, M.S.: (Un)Real animals: anthropomorphism and early learning about animals. Child Dev. Perspect. **10**(1), 10–14 (2016)
61. German, D., Sterk, C.E.: Looking beyond stereotypes: exploring variations among crack smokers. J. Psychoactive Drugs **34**(4), 383–392 (2002)
62. Gogoll, J., Müller, J.F.: Autonomous cars. In favor of a mandatory ethics setting. Sci. Eng. Ethics **23**(3), 681–700 (2017)
63. Goodall, N.J.: Machine ethics and automated vehicle. In: Meyer, G., Beiker, S. (eds.) Road Vehicle Automation, pp. 93–102. Springer International Publishing (2014)
64. Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., Ditto, P.H.: Mapping the moral domain. J. Pers. Soc. Psychol. **101**(2), 366–385 (2011)
65. Guglielmo, S., Monroe, A.E., Malle, B.F.: At the heart of morality lies folk psychology. Inquiry **52**(5), 449–466 (2009)
66. Guthrie, S.E., Guthrie, S.: Faces in the clouds: A new theory of religion. Oxford University Press on Demand (1995)
67. Habib, M., Cassotti, M., Moutier, S., Houdé, O., Borst, G.: Fear and anger have opposite effects on risk seeking in the gain frame. Front. Psychol. **6**, 253 (2015)
68. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**(4), 814–834 (2001)
69. Haidt, J.: The new synthesis in moral psychology. Science **316**(5827), 998–1002 (2007)
70. Haidt, J., Graham, J., Joseph, C.: Above and below left–right: ideological narratives and moral foundations. Psychol. Inq. **20**(2–3), 110–119 (2009)

71. Hakli, R., Mäkelä, P.: Moral responsibility of robots and hybrid agents. Monist **102**(2), 259–275 (2019)
72. Harari, Y.N.: Homo Deus: A brief history of tomorrow. Random House (2016)
73. Harris, B.J.: The History of the Future: Oculus, Facebook, and the Revolution That Swept Virtual Reality. HarperCollins (2019)
74. Haslam, N.: Dehumanization: an integrative review. Pers. Soc. Psychol. Rev. **10**(3), 252–264 (2006)
75. Haslam, N., Loughnan, S.: Dehumanization and Infrahumanization. Annu. Rev. Psychol. **65**(1), 399–423 (2014)
76. Hastings, G., Stead, M., Webb, J.: Fear appeals in social marketing: Strategic and ethical reasons for concern. Psychol. Mark. **21**(11), 961–986 (2004)
77. Hawken, P. (ed.): Drawdown: The Most Comprehensive Plan Ever Proposed to Reverse Global Warming. Penguin (2017)
78. Hibbard, B.: Avoiding unintended AI behaviors. In: Bach, J., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence, pp. 107–116. Springer, New York (2012)
79. Hildt, E.: Artificial intelligence: does consciousness matter? Front. Psychol. **10**, 1535 (2019)
80. Hirsh, J.B., DeYoung, C.G., Xiaowen, X., Peterson, J.B.: Compassionate liberals and polite conservatives: associations of agreeableness with political ideology and moral values. Pers. Soc. Psychol. Bull. **36**(5), 655–664 (2010)
81. Introna, L.: Maintaining the reversibility of foldings: making the ethics (Politics) of information technology Visible. Ethics Inf. Technol. **9**(1), 11–25 (2007)
82. Jackson, E., Rogers, A.: Salvaging Pascal's Wager. Philos. Christi **21**(1), 59–84 (2019)
83. Jaumotte, F., Lall, S., Papageorgiou, C.: Rising income inequality: technology, or trade and financial globalization? IMF Econ. Rev. **61**(2), 271–309 (2013)
84. Jefferson, A., Bortolotti, L., Kuzmanovic, B.: What is unrealistic optimism? Conscious. Cogn. **50**, 3–11 (2017)
85. Jho, H., Yoon, H.G., Kim, M.: The relationship of science knowledge, attitude and decision making on socio-scientific issues: The case study of students' debates on a nuclear power plant in Korea. Sci. Educ. **23**(5), 1131–1151 (2014)
86. Johnson, D.G., Verdicchio, M.: Why robots should not be treated like animals. Ethics Inf. Technol. **20**(4), 291–301 (2018)
87. Johnson, S.C.: Detecting agents. Philos Trans R Soc Lond Ser B Biol Sci **358**(1431), 549–559 (2003)
88. Kahn, P.H., Reichert, A.L., Gary, H.E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J.H., Gill, B.: The new ontological category hypothesis in human-robot interaction. In: Proceedings of the 6th International Conference on Human-Robot Interaction, 159–160 (2011)
89. Kaku, M.: The Future of Humanity: Terraforming Mars, Interstellar Travel, Immortality, and our Destiny Beyond Earth. Anchor (2018)
90. Kelemen, D.: The scope of teleological thinking in preschool children. Cognition **70**(3), 241–272 (1999)
91. Kelemen, D., Carey, S.: The essence of artifacts: Developing the design stance. In: Margolis, E.E., Laurence, S.E. (eds.) Creations of the mind: Theories of artifacts and their representation, 212–230. Oxford University Press (2007)
92. Kelemen, D., Rosset, E.: The human function compunction: Teleological explanation in adults. Cognition **111**(1), 138–143 (2009)
93. Kelemen, D., Rottman, J., Seston, R.: Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. J. Exp. Psychol. Gen. **142**(4), 1074 (2013)
94. Kellen, D., Klauer, K.C.: Theories of the Wason selection task: a critical assessment of boundaries and benchmarks. Comput. Brain Behav, 1–13 (2019)
95. Koverola, M., Drosinou, M., Palomäki, J., Halonen, J., Kunnari, A., Repo, M., Lehtonen, N., Laakasuo, M.: Moral psychology of sex robots: An experimental study—how pathogen disgust is associated with interhuman sex but not interandroid sex. Paladyn J. Behav. Robo. **11**(1), 233–249 (2020)
96. Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I.R., Košová, M., Kopecký, R., Sundvall, J., & Laakasuo, M.: Non-human superhumans—understanding moral disapproval of neurotechnological enhancement https://psyarxiv.com/qgz9c/ (2020, preprint)
97. Kringelbach, M.L., Stark, E.A., Alexander, C., Bornstein, M.H., Stein, A.: On cuteness: Unlocking the parental brain and beyond. Trends Cog. Sci. **20**(7), 545–558 (2016)
98. Kunnari, A., Sundvall, J. R., & Laakasuo, M. (2020). Challenges in Process Dissociation Measures for Moral Cognition. Frontiers in Psychology, 11
99. Laakasuo, M., Sundvall, J., Drosinou, M.: Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. Sci. Rep. **7**(1), 1–10 (2017)
100. Laakasuo, M., Drosinou, M., Koverola, M., Kunnari, A., Halonen, J., Lehtonen, N., Palomäki, J.: What makes people approve or condemn mind upload technology? Untangling the effects of sexual disgust, purity and science fiction familiarity. Palgrave Commun. **4**(1), 1–14 (2018)
101. Laakasuo, M., Palomäki, J., Köbis, N.: Moral uncanny valley: a robot's appearance moderates how its decisions are judged. Int. J. Soc. Robot., 1–10 (2021)
102. Laakasuo, M., Repo, M., Berg, A., Drosinou, M., Kunnari, A., Koverola, M., Saikkonen, T., Hannikainen, I. R., Visala, A. & Sundvall, J. (2021b). The Dark Path to Eternal Life: Machiavellianism Predicts Approval of Mind Upload Technology. *Personality and Individual Differences*
103. Laakasuo, M., Köbis, N., Palomäki, J., Jokela, M.: Money for microbes—Pathogen avoidance and out-group helping behaviour. Int. J. Psychol. **53**, 1–10 (2018)
104. Laakasuo, M., Sundvall, J., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., Koverola, M., Repo, M., Saikkonen, T., Palomäki, J.: Moral psychology andartificial agents (Part 1): the Transhuman connection. In: Thompson, S. (ed.) Machine Law, Ethics and Morality in the Age of Artificial Intelligence IGIGlobal, Steven New York. Retrieved February 2, 2021, from http://moim.fi/MoralPsychologyAndArtificialAgents_Part1.pdf (2021)
105. Laakasuo, M., Sundvall, J., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., Koverola, M., Repo, M., Saikkonen, T., Palomäki, J.: Moral psychology andartificial agents (Part 2): the Transhuman connection. In: Thompson, S. (ed.) Machine Law, Ethics and Morality in the Age of Artificial Intelligence IGIGlobal, Steven New York. Retrieved February 2, 2021, from http://moim.fi/MoralPsychologyAndArtificialAgents_Part2.pdf (2021)
106. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. **10**, 1096 (2019)
107. Larsson, S., Heintz, F.: Transparency in artificial intelligence. Internet Policy Rev. **9**(2), 1–16 (2020)
108. Lawson, E. T., & McCauley, R. N. (1990). Interpretation and explanation: Problems and promise in the study of religion. Religion and Cognition: aReader, 12–35
109. Lawson, R.P., Mathys, C., Rees, G.: Adults with autism overestimate the volatility of the sensory environment. Nature Neurosci. **20**(9), 1293 (2017)
110. Lee, M.K.: Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc. **5**(1), 2053951718756684 (2018)
111. Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Mind. Mach. **17**(4), 391–444 (2007)

112. Light, M.T., Massoglia, M., King, R.D.: Citizenship and punishment: the salience of national membership in US criminal courts. Am. Sociol. Rev. **79**(5), 825–847 (2014)
113. Ludeke, S., Johnson, W., Bouchard, T.J.: Obedience to Traditional Authority:" A heritable factor underlying authoritarianism, conservatism and religiousness. Personality Individ. Differ. **55**(4), 375–380 (2013)
114. MacIntyre, A.: Dependent Rational Animals: Why Human Beings Need the Virtues. Open Court, Chicago (1999)
115. Malle, B.F., Magar, S.T., Scheutz, M.: AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In: Ferreira, M.I.A., et al. (eds.) Robotics and well-being, pp. 111–133. Springer, Cham (2019)
116. Marcus, G., Davis, E.: Rebooting AI. Vintage, New York (2020)
117. Martin, J.W., Young, L., McAuliffe, K.: The impact of group membership on punishment versus partner choice. **(in press)**
118. Martin, L.H., Wiebe, D.: Pro-and assortative-sociality in the formation and maintenance of religious groups. In: Martin, L.H., Wiebe, D. (eds.) Conversations and Controversies in the Scientific Study of Religion, pp. 129–142. Brill (2016)
119. Matthias, A.: Robot lies in health care: when is deception morally permissible? Kennedy Inst. Ethics J. **25**(2), 169–162 (2015)
120. McCann, S.: Conservatism, openness, and creativity: patents granted to residents of American States. Creat. Res. J. **23**(4), 339–345 (2011)
121. McIntyre A.: Doctrine of Double Effect. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Retreived February 2, 2021, from https://stanford.library.sydney.edu.au/entries/double-effect/ (2004)
122. Meacham, D., Studley, M.: Could a robot care? It's all in the movement. In: Lin, P., Abney, K., Jenkins, R. (eds.) Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, pp. 98–111. Oxford University Press, New York (2017)
123. Melson, G.F., Kahn, P.H., Beck, A., Friedman, B., Robert, T., Garrett, E., Gill, B.T.: Children's behavior toward and understanding of robotic and living dogs. J. Appl. Dev. Psychol. **30**(2), 92–102 (2009)
124. Melson, G.F., Kahn, P.H.K., Beck, A., Friedman, B.: robotic pets in human lives: implications for the human-animal bond and for human relationships with personified technologies. J. Soc. Issues **65**(3), 545–567 (2009)
125. Mercier, H., Sperber, D.: The enigma of reason. Harvard University Press (2017)
126. Mikhail, J.: Universal moral grammar: Theory, evidence and the future. Trends Cog. Sci. **11**(4), 143–152 (2007)
127. Mitchell, M.: Artificial Intelligence—A Guide for Thinking Humans. Pelican, New York (2019)
128. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcementlearning. Retrieved February 2, 2021. arXiv preprint arXiv:1312.5602
129. Mohr, J., Sengupta, S., Slater, S.: Marketing of high-technology products and innovations, 3rd edn. Prentice Hall, Upper Saddle River, NJ (2010)
130. Monton, B.: How to avoid maximizing expected utility (2019)
131. Moutier, S., Angeard, N., Houde, O.: Deductive reasoning and matching-bias inhibition training: evidence from a debiasing paradigm. Think. Reason. **8**(3), 205–224 (2002)
132. Müller, C.P., Schumann, G.: Drugs as instruments: a new framework for non-addictive psychoactive drug use. Behav. Brain Sci. **34**(6), 293–310 (2011)
133. Napier, J.L., Luguri, J.B.: Moral mind-sets: abstract thinking increases a preference for "Individualizing" over "Binding" Moral foundations. Soc. Psychol. Pers. Sci. **4**(6), 754–759 (2013)
134. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 427–436 (2015)
135. Nicholson, N., Soane, E., Fenton-O'Creevy, M., Willman, P.: Personality and domain-specific risk taking. J. Risk Res. **8**(2), 157–176 (2005)
136. Nielsen, J.: The Distribution of Users' Computer Skills: Worse than You Think, p. 13. Nielsen Norman Group (2016)
137. Nunn, K.B.: Race, crime and the pool of surplus criminality: or why the war on drugs was a war on blacks. J. Gender Race Just. **6**, 381 (2002)
138. Oldfield, F., Barnosky, A.D., Dearing, J., Fischer-Kowalski, M., McNeill, J., Steffen, W., Zalasiewicz, J.: The Anthropocene review: its significance, implications and the rationale for a new transdisciplinary journal. Anthropocene Rev. **1**(1), 3–7 (2014)
139. Omohundro, Stephen M.: In: Wang, P., Goertzel, B., Franklin, S. (eds.) The Basic AI Drives. IOS, Amsterdam (2008)
140. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books (2016)
141. Palazzo, G., Krings, F., Hoffrage, U.: Ethical blindness. J. Bus. Ethics **109**(3), 323–338 (2012)
142. Parthemore, J., Whitby, B.: Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. Int. J. Mach. Conscious **6**(02), 141–161 (2014)
143. P Pearce, D.: Non-materialist physicalism: an experimentally testable conjecture. Retrieved February 2, 2021, from https://www.hedweb.com/physicalism/ (2016)
144. Penrose, R.: Mechanisms, microtubules and the mind. J. Conscious. Stud. **1**(2), 241–249 (1994)
145. Pinker, S.: Enlightenment Now: The Case for Reason, Science, Humanism, and Progress. Penguin (2018)
146. Preston, J.L.: The egocentric teleological bias: how self-serving morality shapes perceptions of intelligent design. In: Gray, K., Graham, J. (eds.) Atlas of Moral Psychology, pp. 352–359. The Guilford Press, New York (2018)
147. Putt, S.S., Wijeakumar, S., Franciscus, R.G., Spencer, J.P.: The functional brain networks that underlie early stone age tool manufacture. Nat. Hum. Behav. **1**(6), 0102 (2017)
148. Rabin, M., Vayanos, D.: The Gambler's and hot-hand fallacies: theory and applications. Rev. Econ. Stud. **77**(2), 730–778 (2010)
149. Riek, L.D., Rabinowitch, T., Chakrabarti, B., Robinson, P.: How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, 245–246 (2009)
150. Kavussanu, M., Roberts, G.C., Ntoumanis, N.: Contextual influences on moral functioning of college basketball players. Sport Psycholog. **16**(4), 347–367 (2002)
151. Rode, C., Cosmides, L., Hell, W., Tooby, J.: When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. Cognition **72**(3), 269–304 (1999)
152. Rothbart, D., Barlett, T.: Rwandan radio broadcasts and hutu/tutsi positioning. In: Moghaddam, F., Harré, R., Lee, N. (eds.) Global Conflict Resolution Through Positioning Analysis, pp. 227–246. Springer Science & Business Media (2008)
153. Schäffner, V.: Caught up in ethical dilemmas: an adapted consequentialist perspective on self-driving vehicles. In: Robophilosophy/TRANSOR, pp 327–335 (2018)
154. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., ergus, R.: Intriguing properties of neural networks. arXiv:1312.6199 (2013)
155. Selby, E.C., Shaw, E.J., Houtz, J.C.: The creative personality. Gifted Child Q. **49**(4), 300–314 (2005)

156. Sessa, B.: The psychedelic renaissance: Reassessing the role of psychedelic drugs in 21st century psychiatry and society, 2nd edn. Muswell Hill Press (2017)
157. Sharkey, A., Sharkey, N.: Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf. Technol. **14**(1), 27–40 (2012)
158. Shim, J., Arkin, R.C.: A taxonomy of robot deception and its benefits in HRI. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2328–2335 (2013)
159. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van der Driessche, G., Hassabis, D.: Mastering the game of go without human knowledge. Nature **550**(7676), 354–359 (2017)
160. Sinnott-Armstrong, W.: "Consequentialism", The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Zalta, E.N. (ed.), Retrieved February 2, 2021, from https://plato.stanford.edu/archives/sum2019/entries/consequentialism/ (2019)
161. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal. Int. J. **24**(2), 311–322 (2004)
162. Sober, E., Wilson, D.S.: Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press (1999)
163. Sperber, D., Mercier, H.: Why a modular approach to reason? Mind Lang. **33**(5), 533–541 (2018)
164. Statt, N.: DeepMind's StarCraft 2 AI is Now Better Than 998 Percent of All Human Players. (2019). Retrieved on February 2, 2021
165. Steinhoff, U.: The secret to the success of the doctrine of double effect (and Related Principles): biased framing, inadequate methodology, and clever distractions. J. Ethics **22**(3–4), 235–263 (2018)
166. Tan, X.Z., Vázquez, M., Carter, E.J., Morales, C. G., Steinfeld, A.: Inducing bystander interventions during robot abuse with social mechanisms. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 169–177) (2018)
167. Tegmark, M.: Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf (2017)
168. Thellman, S., Silvervarg, A., Ziemke, T.: Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. Front. Psychol. **8**, 1962 (2017)
169. Tomasello, M.: Becoming Human: A Theory of Ontogeny. Belknap Press (2019)
170. Tononi, G.: PHI: A Voyage from the Brain to the Soul. Pantheon Books (2012)
171. Tooby, J., Cosmides, L.: Conceptual foundations of evolutionary psychology. In: Buss, D.M. (ed.) The Handbook of Evolutionary Psychology, pp. 5–67. Wiley, Hoboken (2005)
172. van Leeuwen, F., Dukes, A., Tybur, J.M., Park, J.H.: Disgust sensitivity relates to moral foundations independent of political ideology. Evol. Behav. Sci. **11**(1), 92–98 (2017)
173. Varoufakis, Y.: And the Weak Suffer What They Must? Europe, Austerity and the Threat to Global Stability. Random House (2016)
174. Varoufakis, Y.: Adults in the Room: My Battle with Europe's Deep Establishment. Random House (2017)
175. Voci, A.: The link between identification and in-group favouritism: effects of threat to social identity and trust-related emotions. Br. J. Soc. Psychol. **45**(2), 265–284 (2006)
176. Wachsmuth, I.: Robots like me: challenges and ethical issues in aged care. Front. Psychol. **9**, 432 (2018)
177. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right From Wrong. Oxford University Press (2008)
178. Ward, A.F., Olsen, A.S., Wegner, D.M.: The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead. Psychol. Sci. **24**(8), 1437–1445 (2013)
179. Warwick, K.: Artificial Intelligence: The Basics. Routledge, London (2013)
180. Waytz, A., Epley, N., Cacioppo, J.T.: Social cognition unbound: insights into anthropomorphism and dehumanization. Curr. Dir. Psychol. Sci. **19**(1), 58–62 (2010)
181. Waytz, A., Heafner, J., Epley, N.: The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. **52**, 113–117 (2014)
182. West, G.B.: Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. Penguin (2017)
183. Whitby, B.: Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. Interact. Comput. **20**(3), 326–333 (2008)
184. Whitehouse, H., Martin, L.H. (eds.): Theorizing religions past: Archaeology, history, and cognition., Rowman Altamira (2004)
185. Wilks, J., Austin, D.A.: Evaluation of a strategy for changing group stereotypes of the heroin user. Drug Alcohol Rev. **10**(2), 107–113 (1991)
186. Zarouali, B., Dobber, T., De Pauw, G., de Vreese, C.: Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media. Commun. Res., 0093650220961965 (2020)
187. Zeigler-Hill, V., Noser, A.E., Roof, C., Vonk, J., Marcus, D.K.: Spitefulness and moral values. Personality Indiv. Diff. **77**, 86–90 (2015)