ARTICLE

<https://doi.org/10.1057/s41599-022-01060-4>

OPEN

Incompleteness of moral choice and evolution towards fully autonomous AI

Tomas Hauer¹✉

Nowadays, it is fashionable to add the attribute “with artificial intelligence” to all possible devices, platforms and machines. The problem of ethical decision-making, viewed from the perspective of computer, technical and natural sciences, lies only in the complexity of the topic. AI scientists and developers basically proceed from the Turing machine model, assuming that a machine can be constructed to resolve any problems (including ethical decision-making issues) that can mechanically calculate a particular function if this function can be put into an algorithm. Thus, ethical decision-making is conceived as an abstract concept whose manifestation does not depend on the particular physical organism in which the algorithm takes place, nor on what it is made of. Whether from photons, mechanical relays, quantum fluctuations, artificial neurons or human nerve cells. If in practice, a sufficiently complex algorithm is built, it will also exhibit sufficiently complex behavior that can be characterized as ethical in the full sense of the word. This article reflects the main argument that if a task requires some form of moral authority when it is performed by humans, its full automation, transferring the same task to autonomous machines, platforms, and AI algorithms, necessarily implies the transfer of moral competence. The question of what this competence should include presupposes empirical research and reassessing purely normative approaches in AI ethics.

¹Department of Philosophy, Faculty of Philosophy and Arts, Trnava University in Trnava, Hornopotocna street 23, 918 43 Trnava, Slovakia.
✉email: tomas.hauer@truni.sk

Evolution towards fully autonomous AI platforms

The basic thesis of the supporters of the concept—artificial moral agents (AMAs), is that the whole spiritual and mental sphere, including ethics, can be described and understood functionally. (Allen et al., 2000; Allen and Wallach, 2011; Anderson and Anderson, 2010), i.e., as a complex function with a specific program that can be physically realized in any way, similarly to the same program being realized in two digital platforms with a totally different architecture or flying function in a different way and in such different facilities as hummingbird and drone. How to decide what steps are morally right is one of the most difficult questions in our lives. Understanding the ethical pitfalls and challenges associated with these decisions is vital to building increasingly autonomous AI platforms (Allen et al., 2005; Allen et al., 2006; Anderson and Anderson, 2007; Boden, 2016; Boddington, 2017).

Autonomous weapons systems (AWS) are an empirically investigable, extreme and ethically controversial example of such decision-making (Beck et al., 2016). These are weapon systems that would be able to decide on the use of lethal force without human involvement in the decision loop (human-out-of-the-loop). Fully autonomous weapons do not yet exist, but technology is evolving towards their development, and there are already autonomous robots and AI platforms that we could be considered their precursor. As the Human Rights Watch report points out:

“The examples described in this report show that a number of countries, most notably the United States, are coming close to producing the technology to make complete autonomy for robots a reality and have a strong interest in achieving this goal” (HRW, 2012, p. 3).

The autonomy of the weapon system assumes the ability of the machine to perform actions without human control based on interactions between configured computer programs that activate its physical components in the context of the environment (Williams and Scharre, 2015). Today, three types of weapons are distinguished according to the degree of autonomy: human-in-the-loop, human-on-the-loop a human-out-of-the-loop (HRW, 2012, p. 2). In the first category of human-in-the-loop, the aim of system autonomy is to reduce the operator load and increase overall efficiency. Such a robot may not be able to perform tasks on its own without human intervention, but may perform sub-functions, prioritize, track and select targets autonomously, while man has to decide on target hit (Leeper et al., 2012). The human-on-the-loop system is capable of tracking and hitting targets without a human operator, but the actions of such a robot are reversible and supervised, and autonomous mode is only involved in situations requiring a rapid response that man is unable to provide. Weapon systems are usually used to protect ships, ground equipment or vehicles against incoming fire (Boulainin and Verbruggen, 2017). Human-out-of-the-loop or fully autonomous weapon systems are also called robot killers; they can perform actions independently without interaction with the human operator. Fully autonomous weapon systems are deadly devices designed to identify enemy targets that are selected using sophisticated software. Such autonomous systems include a mobile combat platform, such as crewless aerial vehicles, ships or land vehicles, environmental sensors, systems for tracking and classifying objects of interest, and algorithms directing the platform to launch attacks (Galliot and Lotz, 2016). The human-on-the-loop and human-out-of-the-loop systems have similar functions since, in the first case, the autonomous mode can be switched on, in which the machine will act as a fully autonomous weapon.

Researchers in the area (AWS) and AI ethics often point out that the claim that a robot is autonomous tends to reflect only the

fact that it is able to act independently of immediate human control, but still within human-made actions and decisions.

“Sometimes the claim that a weapon is ‘autonomous’ means only that it is capable of acting independently of immediate human control; typically it means that it is a ‘fire and forget’ system capable of determining its trajectory or pursuing its target to some limited extent. Many existing autonomous weapon systems are of this type. Weapons of this nature do not in themselves raise any ethical questions beyond those raised by other modern long range weapons” (Sparrow, 2007, p. 65).

However, this does not necessarily mean that the next generation of intelligent robots will not be able to act and decide on attacks independently in a so-called intelligent way. The actions of these machines will be based on certain reasons, but these are generated by the system itself without human intervention. Owing to the increasing intelligence of robots, their ability to analyze previous experiences improves. In practice, this means that the operation of these machines can become unpredictable, which opens the door to difficult ethical issues regarding their use (Sparrow, 2007). Robot killers, or fully autonomous weapons, would be able to decide on the use of deadly force without human involvement in the decision-making loop (human-out-of-the-loop). Once a fully autonomous system is activated, it can track, select and destroy targets without subsequent human command and control. Autonomy in a broader sense means the ability to make independent decisions, or the ability to act without the participation of outside participants. From a technical point of view, autonomy allows the system to perceive the surrounding environment and direct its activities towards a specific goal. Out-of-the-loop weapons (also called autonomous weapons) are robots that are able to select targets and deliver power without any human input or interaction, as well as robotic AI weapons that can search, identify, select, and attack targets without real-time human control. Such weapon systems can be described as “automated” if their ability to detect and attack targets autonomously is restricted to a relatively limited, predefined and controlled environment. However, if they are able to perform these tasks independently in an open and unpredictable environment, they are referred to as “fully autonomous” (Galliot and Lotz, 2016). Achieving such a level of autonomy raises discussions in the international community about the legitimacy of further robot autonomy through artificial intelligence that allows intelligent machine behavior, and about the ethical use of these platforms. For the purposes of this text, I will further focus on human-out-of-the-loop systems, and on a number of critical arguments that arise in relation to their deployment. In an analysis of the typical argumentation strategies used in discourse about (AWS), I will try to show that their authors tacitly assume a certain role that ethics should play in this context and non-reflectively adopt settled stereotypes about what we should expect from AI ethics and in what specific way it can be useful for us.

Autonomy: the power to decide

The first argumentation strategy against human-out-of-the-loop systems most often emphasizes the assertion that fully autonomous weapons would undermine human dignity, because as inanimate machines, they cannot understand or respect the value of life, yet would have the power to decide when and under what circumstances they can deprive somebody of it (Goose, 2017). This argument says that completely autonomous weapons could undermine the principle of dignity, which lies at the heart of international human rights, claiming that every person is worthy of respect. A lifeless machine cannot truly respect the value of human life or understand the meaning of its loss. Allowing the machine to make decisions about when to take a life would

fundamentally undermine the meaning associated with such a decision and damage the very concept of human dignity (Docherty, 2014; HRW, 2018). These sentences emphasize the idea that it is an insult to the dignity of an individual if a decision is made to kill them with a machine that can never understand what the value of human life means. It would not be an insult to human dignity if robots had power over the life and death of humans (Heyns, 2017). The common denominator is the assumption that the machine cannot be subject to moral authority, but can only follow the values of its programmers. For example, Asaro is of the opinion that it is necessary to maintain the superiority of human morality, law and justice. Otherwise, we would abandon the concept of human dignity and degrade man to a mere object (Asaro, 2012). All arguments are shared by the assumption that human dignity is an unconditional intrinsic value that has its source in human autonomy. According to this view, the existence of AWS affects a person's ability to decide his or her choices and, therefore, is an attack on human dignity.

The second argumentation strategy highlights the problem of accountability for decisions. A weapon or robot cannot be held responsible for its actions, and if unnecessary civilian deaths or casualties occur, it is unclear who could be punished or charged (Asaro, 2012; Tamburrini, 2016). Other arguments point to a lack of human judgment, common sense, understanding of human intentions and values. Decisions in the case of ethical dilemmas require human judgment because the "rules" that make them require an interpretation that is completely and fundamentally different from explicit rules such as chess. The very nature of moral competence assumes that warriors will be human agents (Asaro, 2012). In order to become moral agents, AWS must demonstrate the ability of empathic interest in a man and a true understanding of human social behavior. Heyns also asserts that the law generally requires human judgment; justice requires a human duty to consider evidence, deliberate alternative interpretations, and obtain an informed opinion. The structure of law and justice processes require human presence (Heyns, 2017). The absence of human reasoning would make any decisions about losing life arbitrary and irresponsible.

Finally, the third argumentation strategy refers to the assumption that AI can never, in principle, cross a clear boundary, a line indicating a fundamental ontological difference between humans and any machine that could ever be constructed (Bryson, 2008; Sharkey and Sharkey, 2011; Tonkens, 2009). When asked whether machines, robots, and advanced AI algorithms could become fully ethically autonomous in the near future, most philosophers and ethics answer that no, because AI has no free will and is unable to realize phenomenal consciousness (Nagel, 1974; Chalmers, 1995; Chalmers, 1996, p. 93), which are absolutely necessary prerequisites for the exercise of moral authority. This rhetorical figure emphasizes that we must not allow a situation in which it would be very difficult to ascertain whether advanced AI algorithms made a mistake or reacted to something beyond human perception. The machine is only able to make independent decisions if it has encountered such a situation in the past and has already set up an algorithm of operation for a particular case, but still cannot generalize previous experience and adapt independently in an unknown environment (Endsley, 2017). In addition, the incompleteness and uncertainty that may occur in the first stages of information processing will further spread and deepen the error. According to critics, the possibility of independent decision-making increases the unpredictability of systems and the risk of choosing a completely random plan. (Ansell, 2014). Machine perception error rate, framing problem (Wallach, 2010; Deng, 2015), machine learning imperfections, inability to consider different scenarios, the limited possibility of independent decision-making, etc. are all "minor" difficulties and

problems in contrast to the lack of free will and absence of consciousness in the case of machines. The concept of phenomenal consciousness is intended to capture those aspects of experience that are objectively indescribable. I cannot tell others how exactly I experience, see, feel the world. One reason why these experiences are indescribable is that they are inward, or inseparable from our subjectivity and, therefore, inaccessible to the perspective of a third party (Dennett, 1988, p. 358). The objective description concerns facts that are public, equally accessible to all. Nevertheless, phenomenal qualities are intrinsically private, available only to us. Finally, while the objective description may be more or less correct and can be further reviewed, the phenomenal characteristics of experience are, as they are immediately apparent to the subject. In the contemporary philosophy of mind, the term "qualia" has become established for the qualities of experience that share the four characteristics mentioned above—indescribability, inwardness, privateness, and immediate obviousness.

Kant's moral ambivalence and a dispute over the nature of values

The question of whether ethics can be expressed in algorithms depends on how AI developers understand ethics and on the adequacy of their understanding of ethical issues and methodological challenges in this area. There are four basic problem areas with which machine and platform developers containing advanced AI algorithms are confronted—lack of ethical knowledge, pluralism of methods in ethics, cases of ethical dilemmas, and machine bias. Researchers and programmers face at least two types of problems due to their general lack of ethical knowledge. The first type is so-called beginner mistakes, which could be solved by providing these people with at least basic knowledge of ethics. These problems show that researchers and programmers who have made mistakes are not aware of the moral complexity of the issues they are facing or lack the appropriate ethical knowledge. The second, more difficult methodological question concerns areas of mutual disagreement in ethics where there are currently no easy solutions to ethical dilemmas. The question of the degree to which researchers should listen to the views of ethicists and moral philosophers on the sensitive issue of autonomous AI ethics is very complicated. Ethics and moral philosophy often try to convince us that they possess well-founded theories that are based on something completely different from the public's moral ideas. Compared to these ideas, they have a different, much better and more profound source of moral knowledge. The subject of their research is the very nature of moral concepts and the logic of moral arguments. So, if there is a well-founded moral theory, we should be ready to accept the consequences that result from it, even when it forces us to change moral views on serious problems. Ethics and moral philosophy are said to be able to provide such theories and thus correct the lack of our moral competence. In view of all this, it is clear that both ethics and moral philosophy are established philosophical disciplines with traditional canons of texts and experts who interpret these canons and the arguments contained in them and guard their correct interpretations. As a result, the question of "if at all" and possibly "how" to implement ethics in platforms containing advanced AI algorithms are now desperately losing to the question of "what actually". In the next part of the text, I will try to analyze the conceptual and methodological problems associated with the "what actually".

The notion that the concept of the "moral theory" is based on something stronger than a set of moral ideas, intuitions, expectations, prejudices and hopes of people living in a given society is as doubtful as the idea that a moral argument has a kind of special

logic that can only be understood by a person with a specific philosophical education. In my opinion, to explain and grasp a concept is simply to know how we use a certain word. For example, we understand the term “quark” when we know how to talk about quantum mechanics and discuss with potential opponents at the European Organization for Nuclear Research (CERN) in Geneva, we understand the term “allostery” if we can effectively use it in the biochemistry discourse and we write scientific articles on this subject, and the term “cubism” when we are knowledgeable in the history of European painting and are able, as curators, to open an exhibition on this topic, for example, in the Academy of Fine Arts in Florence (Galleria dell’Accademia). However, terms such as—correct, human dignity, value, responsibility, loyalty, respect for authorities, concern for others, etc.—are not professional terms and it is, therefore, not clear at all what kind of special education would make one understand the meaning of these words better than writers, lawyers, artists, psychologists, or the public. Moreover, it will be very difficult to find any meaning of the word “logic” where the arguments related to the logic of moral choice will have a different “logic” than those related to the logic of career choice, consideration of stock purchases or voting in democratic elections. Ethics and moral philosophy do not formulate almost any reason to believe that it is through philosophical education and not, for example, education in the history of art, sociology, cognitive psychology or law, that we should gain a better ability to govern the logic of moral decision-making. Therefore, let us try to propose an alternative explanation for how the arguments from many academic articles, books and scientific studies differ in current ethics and moral philosophy from that of the lay public.

Perhaps the answer to this question might be that ethicist may have no deeper insight than non-specialists, but are more willing to take some of the views of the German philosopher Immanuel Kant (Kant, 1959; Kant, 1983; Kant, 1993; Kant, 1994) seriously. Compared to other philosophers, Kant used terms such as “the nature of moral concepts” and “the logic of moral argument” more frequently and with great respect. In fact, he claimed that morality is not like anything else in the world, that it is quite unique. Kant was of the opinion that there was a huge and insurmountable difference between two areas—the area of rationality and the area of morality. If you agree with him, like many moral philosophers (Kerstein, 2002; Wood, 1999; Denis, 2010), you are close to believing that studying the essence of moral concepts can be transformed into a scientific discipline. However, if you have not read Kant, or if reading the *Foundation of the Metaphysics of Morals* have not engaged your attention, the idea that morality may be the subject of scientific research may seem strange. Nevertheless, if, for various reasons, you take Kant’s ethics seriously, an approach to moral research based on the assumption that there is a particular source of moral principles may seem credible (Reath, 2006). However, you may think that moral principles are nothing more than a certain range of our moral intuitions. The principles are good for summarizing a range of moral responses, but they do not have the power to correct them. The principles draw all their strength and persuasiveness from our moral intuitions and prejudices. Let us define intuition as judgments, solutions and ideas that appear in our consciousness without realizing the cognitive process that led to them (Haidt, 2001). Moral intuitions are their specific subset related to the feeling that I approve or reject something. Thus, moral intuitions are such heuristic cognitive abbreviations. We follow them because we usually do not have enough time to look for other ways and think about their correctness.

Kant, on the one hand, helped us get rid of the notion that morality is a matter of God’s commandment, but, unfortunately, he also retained the idea that morality was about unconditional

obligations. The conclusions drawn in many texts by representatives of contemporary ethics and moral philosophy often convince us that Kant made a fascinating discovery and gave us a vital new idea, the idea of moral autonomy. In the sense of obeying the unconditional command of reason, moral autonomy is a very special professional term—which needs to be learned in a similar way as any other professional term, i.e., by mastering a very specialized language game called deontological ethics. This language game must be mastered by anyone wishing to obtain a doctorate in moral philosophy, although many people who have made difficult moral decisions in various situations throughout their lives can survive without it quite well. A large number of texts in AI ethics today take for granted a discourse in which the idea that “morality” still denotes a rather mysterious entity that requires intensive study is not questioned. Reading Aristotle (Anagnostopoulos, 1994; Bostock, 2000), Milla (Crisp, 1997; Becker, 2004), or Kant (Wood, 1999) is a good way to be initiated into this discourse. Kant’s critics, on the other hand, hoped that the way Kant talked about moral decisions would eventually become appealing to fewer and fewer people. We will consider separating morality from rationality a very bad idea, and we will especially refuse the idea that the moral imperative has a source other than the sensible advice of a friend what to do. Therefore, I propose that Kant be understood as a thinker whose moral view could never be reconciled with Darwin’s naturalistic description of the origin of man. Every exploration research—in ethics, physics, cognitive science, art, politics or logic—is a matter of recontextualization. You can see physics as the way we try to cope with some aspects of the universe, philosophy, or art as ways that help us cope with other aspects of the universe. Some investigations result in assertions, graphs, theorems or mathematical models; others result in images, categorical imperatives or stories. The question of what claims to hold and which stories to tell is a question of what will help us achieve what we want or what we should want. All our judgments are experimental and inaccurate in this examination. Unconditional and absolute are not things that we should pursue in ethics.

In my opinion, one of the reasons for the lack of clarity and conceptual confusion in current research on ethically autonomous AI is that we have got stuck in the trap Kant had set for us. Most AI researchers are convinced naturalists who would like to reconcile their moral views with Darwin’s view of human origin. Naturalists cannot accept Kant’s idea of a particular moral motivation or the idea of a capability called “reason” that commands mankind without the considerable conceptual problems that this raises. For naturalists, the assertion that moral principles have no intrinsic nature is to say that they do not come from any particular source. They simply grow out of our clashes with the environment, just as hypotheses about planetary motion, criminal law, quantum mechanics, etiquette, democratic politics and many other types of our linguistic behavior are born. And like them, ethical principles are good only to the extent that they lead to good consequences, not because they are in a special relationship to either part of the universe or the human mind.

Therefore, I suggest that in the discussion of AI ethics we get rid of the notion that moral philosophy is obliged to equip us with moral principles, which are absolutely independent of the specific context in the sense that they can be applied to any situation. The fact that a given principle best addresses a moral dilemma does not imply that it must be the best in every context. If we assume this, it is similar to the assumption that any decision and rule agreed by a courting couple that is just getting married depends on the authority of the divorce court, because that court will have the final say in resolving all their disputes unless they are able to solve them otherwise... and if they ever get divorced. Let us provisionally call any principle established by the above

argumentation Kant's moral ambulance. We do not need and do not use such an ambulance everyday or every month, but it receives and gains absolute priority in emergency situations. These are cases and emergencies such as suddenly having a concrete reason to criticize some of our unquestionable moral beliefs or face new problems and challenges, or getting in contact with people whose morality or culture we do not know and understand. It is when professors of moral philosophy convince us that we need Kantian formulations of a categorical imperative. Similarly, in this situation, we could use the principle of utilitarianism, according to Stuart Mill (Becker, 2004). Both types of principles are unlimited in general, which can help us reach a reasonable agreement in certain negotiation and decision-making situations where our too general or too specific justifications cease to work. Kant's categorical imperative only recommends that we always ask ourselves how we would like us to be treated in a similar situation. An attempt to achieve something more by it, to get ready rules to solve any moral problem immediately, should be considered an attempt born of fear and nourished by the love of authority. Only admiration for authority can lead us to the idea that the absence of fixed, firm and universally applicable ready-made principles equals moral chaos.

Deep-rooted methodological problems

Although many authors in AI ethics now think that Mill's or Aristotle's ambulance is better than Kant's and are in fiery dispute with opponents over it, in fact, the difference between Mill, Kant, and Aristotle is not particularly interesting. I suppose we should no longer use the assertion that principles are extremely important in our debate on the ethics of artificial moral agents (AMAs), and that if we do not claim allegiance to any of the above emergency service-ambulances (utilitarianism, deontological ethics or ethics of virtue), we are intellectually irresponsible. Deciding which emergency service-ambulance we will conclude a contract with, whether with Kant's, Aristotle's, or Mill's one, is not so important, as we realize that moral principles merely sum up a large number of previous considerations and decisions and that they can only remind us of some of our previous intuitions, prejudices, practices, hopes and expectations. Such abstract remarks in the form of Kant's ambulance can help us only when more specific consideration does not help us resolve disputes with our loved ones. Indeed, Kant's ambulance is not algorithmizable; it only provides guidance. Most of us consider ourselves morally more advanced than the generations of our predecessors. We believe that our moral decisions are much more complex, more sophisticated, more thoroughly premeditated and justified than those of previous generations. At the same time, paradoxically, we do not think we know more general moral truths than our parents. In fact, we are more experienced only in the sense that we have a greater understanding of considering a large number of variants, a better sense of the various alternative scenarios and their possible consequences. In this context, Kantian discourse appears to be outdated, and the idea of researching the "essence of moral concepts and their logic" sounds naive and hopeless.

Kant understands morality as something that arises from a specifically human ability called "reason" and rationality as something we have in common with animals. The origin and development of ethics can be viewed analogously to the origin and development of language. We view the history of language as an uninterrupted story of gradually increasing complexity. Likewise, ethics is a story about how we got from the Neanderthals grunting and nudging elbows, through the Ten Commandments, Pyrrho and Dante, to the German philosophical treatises and the current debate on robots and AI ethics. Both stories are part of a larger story in which biological evolution flows smoothly into

cultural evolution. From an evolutionary point of view, grunting differs from philosophical treatises only in the degree of complexity. According to this view, those moral philosophers who sharply differentiate the reason from experience, or morality from reason try to make a metaphysical difference out of an important micro difference (Greene, 2007). Therefore, they invented problems that are unsolvable and yet fabricated.

Some moral philosophers try to convince us that if we really want to get the right instructions on how we should live or how we should organize society, it is natural to seek advice from thinkers such as Aristotle, Plato, Mill, Kant, Nietzsche, Rawls etc. However, this view raises many questions. In the eighteenth and nineteenth centuries, when religion lost much of its prestige as a source of comfort in a hard human lot, philosophy pushed forward as a substitution. This was the great era of moral philosophy. It was at this time that I. Kant and J. Bentham postulated elegant, simple, abstract principles to summarize and explain all our moral intuition. Their theories wanted to show us how to deal with the torturous moral dilemmas by simple and direct applications of deductive or inductive reasoning. However, very soon after the emergence of deontology and utilitarianism, people quickly began to realize that these principles were just as hopelessly useless as the Ten Commandments or Golden Rule of Behavior. Whatever is morally recognized, it cannot be achieved by memorizing rules or formulating principles. When one is faced with a truly ethical dilemma, the last thing one needs will be general principles.

If we want to understand why scientists and researchers in quantum mechanics, cognitive sciences, mathematics, astronomy or biochemistry agree on a common position or conclusion much more often than literary critics, art historians or moral philosophers, the traditional distinction between objective and subjective, or rational versus irrational will not help us. The uselessness of subjective versus objective dichotomy is also shown in the discussion on the nature of values in AI ethics. Conversely, what can really help us effectively is if we look at the way in which people in different disciplines resolve their disputes. I will further develop my idea of the analogy between research in science and the ethics of AI on the example of the nature of values. I would like to illustrate my position on the dispute over objectivity or subjectivity of values by the following example. One of the evergreen debates on ethics is that statements such as "one should suppress his urge to be cruel to robots and AI platforms, even though they lack consciousness and freedom of choice" cannot be true in the same sense as the sentence that "the content of a square constructed above the hypotenuse of a right-angled triangle is equal to the sum of the contents of the squares above its legs". It is assumed that the statements of "what should be" do not want to inform us about the nature of objects, while the statements of "what is" (Pythagorean theorem) want to inform us about it. In this view, looking for objective values, objects to which the statements of "what should be" would apply—is simply wrong, because statements of this type can be neither true nor false. I believe the statements of "what is to be" in the case of cruelty to robots are as true as the statements of "what is" relating to geometry or gravity. Neither of them expresses the truth about an object. In either of the two examples above, it makes no sense to understand truth as correspondence with reality. No object called "sum of squares or gravity" that correspondence would relate to never existed. Similarly, there is no object called "what it should be" that we should morally respect. Both kinds of statements have a purpose and meaning only as part of an extensive summary of assumptions, concepts, duration, and opinions on how we should live or understand the nature of the universe around us. Both statements (about geometry and avoiding cruelty to robots) are analogous to sentences. There is a number that has

not yet been translated into Dutch, or there is a number whose square root is equal to that number. Such statements are part of some broadly conceived mathematical theory, and mathematicians accept the statements as true because the theory as a whole corresponds to the needs for which theories of such atypical numbers are constructed in mathematics. However, no one imagines that these statements can be compared to some objects. At the same time, it would be naive to say that these mathematical truths are only subjective or only relative. The degree of controversy is not a matter of the metaphysical nature of its subject. Using awkward and vague terms such as “subjective” or “objective” in the case of values is nothing more than a demonstration of cultural backwardness.

Relativity or subjectivity of morality, as opposed to the claims made in science, cannot be proven even by the fact that some authors can promote a very extravagant position. Like the relativity or subjectivity of astronomy, it cannot be proven by the assertion that anyone today can consistently insist that the sun rotates around the earth, or that what Newton called gravity is, in fact, a manifestation of Hegel’s absolute spirit. No one can refute an assertion such as a number higher than 113, or that all but the author of this statement are robots. One can claim many stupid, nonsensical or crazy things about AI ethics that can never be refuted. This is due to the fact that concepts such as deduction, contradiction, argument, logic, etc. are only meaningful in situations where there is at least a minimal match between the competing parties. Thus, we can only contradict or refute the assertion of a person who recognizes at least some of our initial assumptions. We do not consider opponents that have too few assumptions in common with the other parties to the dispute as unbeatable adversaries, but rather as amateurs, too uneducated, naive, or childish to be capable of being involved in a discussion. Each discourse—moral, scientific, mathematical, artistic, legal—has certain entry conditions. The difference between Darwin on the one hand and someone who claims that there has never been a number greater than 113 is that if we accept his original concept, hitherto unknown concepts and paradoxes, we get a better explanation of how the world around us works. If we are to take the scientific theory seriously, it is not enough that it denies the basic principles of another theory, it must offer something more. It must provide us with its own description of the functioning of the thing and show its priority. What does all this mean? It would help us to get rid of not only Kant’s moral ambulances and its siblings in AI research but also the distinction between subjective and objective in terms of values. These distinctions have not made it easier for us to think about morality in the AI context, but rather complicate it.

Naturalism in AI metaethics

“Until the 1980s, however, the vast majority of work in the brain sciences made no references to consciousness. In the last two decades, philosophers, psychologists, cognitive scientists, clinicians, neuroscientists, and even engineers have published dozens of monographs and books aimed at “discovering,” “explaining,” or “reconsidering” consciousness. Much of this literature is either purely speculative or lacks any detailed scientific program for systematically discovering the neuronal basis of consciousness (Koch, 2004, p. 4).”

Naturalism in metaethics corresponds to an approach whereby questions of moral choice are a problem with empirical solutions. All the claims we make about morality and ethical decision-making are claims about the states of the world that we can treat as any other fact of scientific interest. In addition to the normative feature of morality and the related difficulties in applying to the AI area, which I dealt with in the example of Kant’s moral

ambulance, I want to emphasize further even further characteristics of morality, namely that morality is a psychological and social phenomenon. Moral decision-making is established by an individual’s experiencing placed in a particular social situation, which are circumstances that can be analyzed by empirical research. On the one hand, there is the expectation that, since ethics has, historically, researched morality for the longest time, it is able to offer the necessary tools. On the other hand, there is a critical statement that the traditional approach to moral research does not have an empirical but normative base. Philosophical research on morality usually takes the form of non-empirical research of an empirical phenomenon, i.e., research that does not begin with observing how people actually make decisions in their moral lives, but with a normative idea of how they should act. This, however, considerably complicates the ambition to describe morality and moral behavior in the AI context. In general, various kinds and forms of unconscious activities play a major role in resolving ethical dilemmas. Moreover, it is precisely the distortion of moral evaluation that can be empirically analyzed that reveals effects such as an escalation of commitments, reliance on authority, the phenomenon of ethical blindness, the inconsistency of attitude, or reciprocal altruism (Palazzo et al., 2013; Schoemaker and Russo, 2016; Cialdini, 2006).

Retrospective rationalization in moral decision-making

Other examples of retrospective rationalization in moral decision-making show (Haidt, 2001; Haidt, 2012) that in a situation that is unclear and ambiguous, i.e. in a situation where it is not entirely clear to a person what alternatives he can actually choose in his actions, what is right or wrong on the basis of emotionally driven intuition, and then, if necessary, we invent reasons to explain and justify our judgments. Haidt admits that sometimes, some people may properly justify the way to moral conclusions in this way, but insists that this is not a generally accepted norm. More important for our purposes, however, is that Haidt does not distinguish between the different approaches to ethics known to moral philosophers—utilitarianism, deontology, and the ethics of virtue. Rather, his radical thesis is intended to apply to all major branches of contemporary moral philosophy. At the beginning of his book, Haidt then emphasizes that “intuitions come first, strategic reasoning second. Moral intuitions arise automatically and almost instantaneously, long before moral reasoning has a chance to get started, and those first intuitions tend to drive our later reasoning. If you think that moral reasoning is something we do to figure out the truth, you’ll be constantly frustrated by how foolish, biased, and illogical people become when they disagree with you. However, if you think about moral reasoning as a skill we humans evolved to further our social agendas—to justify our own actions and to defend the teams we belong to—then things will make a lot more sense. Keep your eye on the intuitions, and don’t take people’s moral arguments at face value. They’re mostly post hoc constructions made up on the fly, crafted to advance one or more strategic objectives” (Haidt, 2012, p. 11).

What the above-mentioned authors show in their research, in particular by socio-scientific experiments, is confirmed in much more conclusive ways by experiments in neuroscience (Bear et al., 2015; Gazzaniga et al., 2018; Koch, 2004; Ramachandran, 2011). More conclusive because neural imaging truthfully and reliably shows what one feels in a given context—not what one thinks he feels or what he feels he should tell the people around him to strengthen his moral self-image in their eyes. Numerous experiments show that the inclination to make certain moral decisions and take certain moral positions is given to man to a considerable extent by the nature of his organism. Neuroscience research of this kind can reveal what this setting looks like, what is

responsible for it, and what its genesis is. Perceptual neuroscience has advanced to the extent that reasonably sophisticated computational models have been constructed that have proven successful in conducting experimental agendas and summarizing data (Churchland, 2011). The aim of the research is to discover a minimal set of neuronal phenomena and mechanisms together sufficient for specific conscious perception. Neuroscientists have sought the neural substrate of human morality, which creates the biological basis of moral behavior, using various methods such as magnetic resonance or non-invasive methods of brain stimulation. The results show how activity in brain areas associated with affection, cognition, motivation, and other moral, psychological processes is related to moral behavior. These neuroscientific studies of human morality have significantly and fundamentally influenced other traditional moral disciplines, providing a new perspective on the possibilities and ambitions of moral philosophy (Greene et al., 2001; Greene et al., 2004; Greene et al., 2009). There is considerable and growing evidence that suggests that most of what we do is done unconsciously and for reasons that are inaccessible to us (Wilson, 2002). The general sense of research in cognitive sciences seems to lead to the conclusion that our most fundamental moral dispositions are evolutionary adaptations (Bor, 2012), that have arisen in response to the demands and opportunities created by social life. The question why our adaptive moral behavior should be guided by moral emotions and not by anything else, such as moral reasoning, can be answered in this light in such a way that emotions are very reliable, fast and effective responses to recurring situations, while justification is an unreliable, slow and ineffective strategy in everyday moral decision-making. From an evolutionary point of view, it is no surprise that moral dispositions have evolved, nor is it surprising that these dispositions are implemented emotionally in consciousness.

The neuroscientist Christof Koch defines his research program by starting at a level that is most accessible to the public (Koch, 2004, 2012). According to Koch, studying visual perception has several advantages in terms of neuroscience compared to studying other senses, at least with respect to understanding consciousness. First, humans are visual creatures. This is reflected in the large amount of brain tissue devoted to image analysis and the importance of vision in everyday life. If you get the flu, your airways become blocked, and you lose your sense of smell temporarily, but it is not a big limitation. Conversely, the temporary loss of vision, such as during snow blindness in Antarctica, devastates you completely. Second, visual perceptions are vivid and rich in information. Images and films are highly structured but easy to manipulate using computer graphics. Thirdly, visual perception is much more likely to be misleading. This manifests itself in a total infinite number of illusions. Since the end of the 19th century, researchers have been dealing with the issue of sensory illusions, most of which are devoted to optical illusions.

Very often, our senses fail and do not reflect the world correctly. This is particularly evident in the vision because it is through the optical apparatus that we obtain the vast majority of information from external reality. Yet, in controlled conditions, people have enormous difficulties in estimating lengths and sizes. Owing to limited spatial perception, they are unable to properly assess parallelism and perspective, due to the effect of the context, they misinterpret shades and colors, or succumb to the illusion of apparent movement. Analogous, though perhaps less illustrative examples of distorted perception could be found in all other senses. The most eloquent evidence of the limitations of human perception can be considered inattention blindness and change blindness, which is manifested by the inability to notice even very significant events in unclear situations. Our perception is often overwhelmed by incentives that we cannot evaluate effectively

enough, and, therefore, we are indifferent to unusual phenomena. It is for these reasons that Koch's research program focuses on the field of visual perception, deliberately ignoring other important aspects such as intelligence, language or emotions. A similar conclusion was reached by V.S. Ramachandran (Ramachandran, 2003; Ramachandran, 2011) and his utilitarian theory of perception. Visual perception does not involve inferring exactly the kind of rational proofs (nor does it solve complex formulas), nor can we speak of some simple "resonance" to external stimuli ("inputs"). Perhaps a variety of short connections proven by evolution in visual perception are a common strategy in biology. However, visual consciousness is more than other aspects of perception available to empirical research. Similar conclusions can be found in research in the field of synesthesia (Hubbard and Ramachandran, 2005) or, for example, the phenomenon of blindsight, which brings extremely interesting views to our understanding of the issue. For example, suppose the perception of actually a color-blind synesthetic who, when looking at numbers, perceives colors he might not otherwise see? This finding suggests that qualities do not bind to the perceptions themselves, but somehow arise from the processing of information in the brain. In essence, we could describe them as a mark that the brain uses to tag information belonging together (Ramachandran, 2004). Laurence Tancredi also points to the much-needed link between empirical research and morality: "Today, research in evolution, genetics, and neuroscience is showing that what appeared to evolve from the social need of snake in fact far more origins. It now seems more likely that human biology has a certain type of society to be shaped in particular ways. A new science, evolutionary psychology, emerged in the 1990s to focus on explaining human behavior against the backdrop of Darwinian theory. This science considers the biological forces of genetics and neurotransmissions in the brain influence of unconscious strategies and conscious intentions and proposes that these features of biology undergo subtle but continuous change through evolution. Though it is indeed a social construct, morality gets its timelessness and universality from the human brain" (Tancredi, 2005, p. 10). Since both people and AI have only limited ability to reflect the world in its complexity, it is important to focus primarily on research and a better understanding of how decisions are made in different contexts (man versus AI) and how the exercise of moral authority takes incompleteness, inaccuracy or inconsistency of input information or data into account. Everything suggests that even the question of moral choice in the AI context can be seen as a problem with empirical solutions.

Conclusion

"AI researchers are only just beginning to get a handle on how to analyze even the simplest kinds of real decision-making systems, let alone machines intelligent enough to design their own successors. We have work to do" (Russell, 2019, p. 210).

The theme and reasoning contained in Stuart Russell's latest book virtually call for much closer cooperation between AI researchers and academics in the field of moral philosophy. However, the conclusions reached so far correct the idea of what practical ethics as a scientific discipline could offer to AI research. A researcher wishing to know how to behave in a certain type of "everyday" problems related to his research goal will come across a set of impractical and mutually exclusive (yet universal) principles. Not all of the questions raised in this text are original questions for AI ethics in the sense that these or similar questions would not be asked by other disciplines. The possibility of proposing practical recommendations in the field of both AI ethics and AMAs seems to necessarily involve coordinated cooperation with many other disciplines (law, cognitive psychology, bioethics, argumentation theory, etc.). In fact, the issue of

AI ethics involves a number of sub-questions such as—how institutes that implement some moral practice in scientific research are established, how influential the moral reasons for creating some legislation in the AI area are, what social problems are associated with AI moral problems, and what possibilities to address them are known, how moral decision-making takes place and which factors influence them, whether moral statements can be treated as an expression of emotional attitude, whether the argumentation in AI ethics has some fundamental specifics, etc.? However, everything suggests that even the question of moral choice is a problem with an empirical solution. We need completely targeted, not flat-rate measures like Kant's ambulance. These are the main lessons of the current discussions. Only a better version of the current AI ethics is a cure for the highly incomplete and hesitant form of contemporary AI ethics. Tomorrow is never a completely new day, only what has been left off yesterday remains in it.

Data availability

All data analyzed are contained in the paper.

Received: 23 September 2021; Accepted: 19 January 2022;

Published online: 01 February 2022

References

- Allen C, Wallach W, Smit I (2006) Why machine ethics? *IEEE Intell Syst* 21(4):12–17. <https://doi.org/10.1109/MIS.2006.83>
- Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics Inf Technol* 7(3):149–155. <https://doi.org/10.1007/s10676-006-0004-4>
- Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. *J Exp Theor Artif Intell* 12(3):251–261. <https://doi.org/10.1080/0952130050111428>
- Allen C, Wallach W (2011) Moral machines: contradiction in terms of abdication of human responsibility? In: Lin P, Abney K, Bekey GA (eds) *Robot ethics: the ethical and social implications of robotics*. MIT Press, Cambridge, p 55–68
- Anagnostopoulos G (1994) *Aristotle on the goals and exactness of ethics*. University of California Press, Berkeley
- Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. *AI Magazine* 28(4):15–26
- Anderson M, Anderson SL (2010) Robot be good: a call for ethical autonomous machines. *Sci Am* 303(4):15–24
- Ansell D (2014) Research and development of autonomous 'decision-making' systems. In: *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. International Committee of the Red Cross, Geneva
- Asaro P (2012) On banning autonomous lethal systems: Human rights, automation and the dehumanizing of lethal decision-making, special issue on new technologies and warfare. *Int Rev Red Cross* 94(886):687–709
- Bear M, Connors B, Paradiso M (2015) *Neuroscience: Exploring the Brain*, Jones & Bartlett Learning
- Beck S, Bhuta N, Geib R, Liu H, Krebs C (2016) *Autonomous weapons systems: Law, ethics, policy*. Cambridge University Press, Cambridge
- Becker B (2004) *An Introduction to Mill's Utilitarian Ethics*. Cambridge University Press, Cambridge
- Boddington P (2017) *Towards a Code of Ethics for Artificial Intelligence (Artificial Intelligence: Foundations, Theory, and Algorithms)*, 1st edn, Springer
- Boden AM (2016) *AI: Its nature and future*, 1st edn, Oxford University Press
- Bor D (2012) The ravenous brain: how the new science of consciousness explains our insatiable search for meaning. Basic Books, New York, NY
- Bostock D (2000) *Aristotle's ethics*. Oxford University Press, Oxford
- Boulain V, Verbruggen M (2017) Mapping the development of autonomy in weapon systems. Stockholm International Peace Research Institute (SIPRI). <https://www.sipri.org/sites/default/files/Mapping-developmentautonomy-in-weapon-systems.pdf>
- Bryson J (2008) Robots should be slaves. In: Wilks Y (ed.) *Close Engagements with artificial companions: Key social, psychological, ethical and design issue*. John Benjamins Publishing, Amsterdam, p 63–74. Retrieved from Accessed 7 Mar 2017 <https://books.google.nl/books?id=EPznZHeG89cC>
- Chalmers D (1995) Facing up to the problem of consciousness *J Conscious Stud* 2(3):200–219
- Chalmers D (1996) *The conscious mind*. Oxford University Press, Oxford, p 1996
- Churchland PS (2011) *Braintrust: what neuroscience tells us about morality*. Princeton UP, Princeton, NJ
- Cialdini R (2006) *Influence: The psychology of persuasion*, Harper Business; Revised edition
- Crisp R (1997) *Routledge philosophy guidebook to mill on utilitarianism*. Routledge, London and New York, NY
- Deng B (2015) Machine ethics: the robot's dilemma. *Nat News* 523(7558):24. <https://doi.org/10.1038/523024a>
- Dennett D (1988) *Quining Qualia*. In: Marcel A, Bisiach E (eds) *Consciousness in contemporary science*. Oxford University Press, New York, NY
- Denis L (ed.) (2010) *Kant's 'metaphysics of morals': a critical guide*. Cambridge University Press, Cambridge/New York, NY
- Docherty BL (2014) Shaking the foundations: The human rights implications of killer robots. <https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rightsimplicationskillerrobots>
- Endsley M (2017) From here to autonomy: lessons learned from human automation research. *Hum Factor* 59(1):5–27. <https://doi.org/10.1177/0018720816681350>
- Gazzaniga M, Ivry R, Mangun G (2018). *Cognitive Neuroscience: The Biology of the Mind*, W. W. Norton & Company
- Galliot J, Lotz M (2016) Super soldiers: the ethical, legal and social implications, Taylor & Francis. *Emerging technologies, ethics and international affairs*. <https://core.ac.uk/download/pdf/72850344.pdf>
- Goose S (2017) The growing international movement against killer robots. Harvard International Review. Accessed May 17, 2018 from <http://hir.harvard.edu/article/?a=14022>
- Greene JD (2007) The secret joke of Kant's Soul. In: Sinnott-Armstrong W ed. *Moral psychology, Vol. 3: the neuroscience of morality: emotion, disease, and development*. MIT Press, Cambridge, MA, pp. 35–80
- Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD (2004) The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44:389–400. <https://doi.org/10.1016/j.neuron.2004.09.027>
- Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293:2105–2108. <https://doi.org/10.1126/science.1062872>
- Greene JD, Stewart LE, Kelly Lowenberg K, Nystrom LE, Cohen JD (2009) Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* 111:364–371. <https://doi.org/10.1016/j.cognition.2009.02.001>
- Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol Rev* 108(4):814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- Haidt J (2012) *The righteous mind: why good people are divided by politics and religion*. Pantheon, New York, NY
- Heyns C (2017) Autonomous weapons in armed conflict and the right to a dignified life: An African perspective. *South Afr J Hum Right* 33(1):46–71
- Hubbard EM, Ramachandran VS (2005) Neurocognitive mechanisms of synesthesia. *Neuron* 48/3:509–520
- Human Rights Watch (2012) Losing humanity: the case against killer robots. https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf
- Human Rights Watch (2018) Heed the call: a moral and legal imperative to ban killer robots. <https://www.hrw.org/report/2018/08/21/heed-call/moral-and-legal-imperative-ban-killer-robots>
- Kant I (1959) *Foundation of the metaphysics of morals*. BobbsMerrill, Indianapolis
- Kant I (1983) On a supposed right to lie because of philanthropic concerns. Hackett, Indianapolis/Cambridge, pp. 162–166
- Kant I (1993) *Critique of practical reason*, 3rd edn. Prentice-Hall, Upper Saddle River, NJ
- Kant I (1994) *The metaphysics of morals*. In: *Ethical philosophy*. Hackett, Indianapolis, IA
- Kerstein S (2002) *Kant's search for the supreme principle of morality*. Cambridge/New York, NY
- Koch Ch (2004) *The quest for consciousness: a neurobiological approach*. Roberts & Company Publishers
- Koch Ch (2012) *Consciousness: confessions of a romantic reductionist*, Massachusetts Institute of Technology
- Leeper A, Hsiao K, Gossow D, Hsiao K, Ciocarlie M, Takayama L (2012) Strategies for human-in-the-loop robotic grasping [online]. https://www.academia.edu/2733833/Strategies_for_human-in-the-loop_robotic_grasping
- Palazzo G, Krings F, Hoffrage U (2013) Ethical Blindness. *Journal of Business Ethics*, 109 (3), (February 6, 2013). Available at SSRN: <https://ssrn.com/abstract=2212617> or <https://doi.org/10.2139/ssrn.2212617>
- Reath A (2006) *Agency and autonomy in Kant's moral theory*. Oxford University, Oxford/New York, NY
- Ramachandran VS (2003) *The emerging mind*, Profile Books; Main edition
- Ramachandran VS (2004). *A brief tour of human consciousness: from impostor poodles to purple numbers*. Pi Press
- Ramachandran VS (2011) *The tell-tale brain: a neuroscientist's quest for what makes us human*, W. W. Norton & Company
- Russell S (2019) *Human compatible: artificial intelligence and the problem of control*, Viking

- Schoemaker P, Russo JE (2016) Decision-making, The Palgrave Encyclopedia of Strategic Management, https://doi.org/10.1057/978-1-349-94848-2_341-1
- Sharkey N, Sharkey A (2011) The rights and wrongs of robot care. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp. 267–282
- Sparrow R (2007) Killer robots. *J Appl Philos* 24(1):62–77. <https://doi.org/10.1111/j.14685930.2007.00346.x>
- Tamburrini G (2016) On banning autonomous weapons systems: From deontological to wide consequentialist reasons. In: Bhuta N et al. (eds), Autonomous weapons systems: law, ethics, policy. Cambridge University Press, Cambridge, pp 121–141
- Tancredi LR (2005) Hardwired behavior: what neuroscience reveals about morality. Cambridge UP, New York, NY
- Tonkens R (2009) A challenge for machine ethics. *Minds Mach* 19(3):421–438. <https://doi.org/10.1007/s11023-009-9159-1>
- Wallach W (2010) Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics Inf Technol* 12(3):243–250
- Williams A, Scharre P (ed.) (2015) Autonomous systems: issues for defence policymakers. Headquarters Supreme Allied Commander Capability Engineering and Innovation Division, Printed by: NATO Communications and Information Agency
- Wilson TD (2002) Strangers to ourselves: discovering the adaptive unconscious. Harvard University Press, Cambridge, MA
- Wood AW (1999) Kant's ethical thought. Cambridge University Press, Cambridge, New York, NY

Acknowledgements

This work was produced at the Department of Philosophy, Faculty of Philosophy and Arts Trnava University in Trnava and supported by the project - Revised Anthropology: the Concept of the Human in the 21st Century VEGA 1/0174/21.

Competing interests

The author declares no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Tomas Hauer.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022