sciendo

Research Article • DOI: 10.2478/law-2021-0006 LAW • 1 • 2021 • 44–49

Law and Business

Morality in the AI World

Agnieszka Lekka-Kowalik

Institute of Philosophy, John Paul II Catholic University of Lublin, Lublin, Poland

Abstract

AIs' presence in and influence on human life is growing. AIs are seen more and more as autonomously acting agents, which creates a challenge to build ethics into their design. This paper defends the thesis that we need to equip AI with artificial conscience to make them capable of wise judgements. An argument is built in three steps. First, the concept of decision is presented, and second, the Asilomar Principles for AI development are analysed. It is then shown that to meet those principles AI needs the capability of passing moral judgements on right and wrong, of following that judgement, and of passing a meta-judgement on the correctness of a given moral judgement, which is a role of conscience. In classical philosophy, the ability to discover right and wrong and to stick to one's judgement of what is right action in given circumstances is called* practical wisdom*. The conclusion is that we should equip AI with artificial wisdom. Some problems stemming from ascribing moral agency to AIs are also indicated.

Keywords

artificial intelligence • Asilomar Principles • moral judgement • artificial conscience • artificial wisdom

Introduction

The presence of AI in our life is a fact. It is enough to consult the webpage of the Boston Dynamics firm to see what AI robots can do.[1] AI systems are developed to become acting agents and participants in human society. Two recent examples illustrate this fact. The first case caused much debate: according to a March report from the U.N. Panel of Experts on Libya, AI drones attacked human targets without consulting any humans prior to the strike – they autonomously "decided" to start the action.[2] A second case is different but also arresting: a Russian company, Xsolla, fired 150 employees out of a total of 500, relying on the results of an AI's analysis that catalogued the unproductive employees. The decision was made by humans but the reason for the action was provided by an AI and without AI that reason would not be formulated; AI worked here as an expert.[3] Other cases might easily be found.

Having AI systems as agents shaping human life is a real challenge. Once we recognize this fact, we need to adjust our social, legal, and moral structures to meet that challenge. In my paper, I attempt to show in what this challenge consists. I will defend the thesis that we need to equip AIs with artificial conscience and artificial wisdom to make them capable of moral judgements and then able to follow those judgements. I will build an argument in three steps. First, I analyze the classical concept of decision, and second, I will present some of the Asilomar Principles for AI development together with their crucial presuppositions. I will then show that to meet those principles one needs the capability to pass moral judgements on right and wrong, to follow that judgement, and then to formulate the meta-jedgement on the correctness of a given moral judgement, which is a role of conscience. I will also show a few problems stemming from the recognition of AI systems as ethical agents.

1 Boston Dynamics, https://www.bostondynamics.com/ [accessed 9.08.2021].

2 Kyle Mizokami, "For the First Time, Drones Autonomously Attacked Humans. This Is a Turning Point," https://www.popularmechanics.com/military/weapons/a36559508/drones-autonomously-attacked-humans-libya-united-nations-report/ [accessed 9.08.2021].

3 Editor. "Russian Employees Fired: IA Labeled Them Unproductive." https://today.in-24.com/technology/205364.html [accessed 9.08.2021].

† Corresponding author: Agnieszka Lekka-Kowalik

E-mail: alekka@kul.pl

Decision as self-determination

In what follows I appeal mainly to an understanding of decisions developed by Stanisław Kamiński and Mieczysław A. Krąpiec[4] within the tradition of so-called classical philosophy.[5] According to this view, any decision is an act of making a theoretical-practical judgement "this should be done" (provided by intellect) and a practical-practical judgement: "I will do this" (determined by will). So any decision is a result of the interplay of intellect and will, which mutually influence each other. The decision is a concrete and unique judgement that cannot ever be repeated in the same circumstances. Seen in that way, the decision is not a rule-following – it involves creativity as a synthesis of knowledge and experience. One decides to do something when a certain state of affairs to be brought about by action specified in a decision is seen as good and desirable. Let us also stress that the decision as a practical judgement does not follow directly and deductively from a general principle. A simple commonsense case might serve as an example: the principle of "feed the hungry" is valid, but from it, one cannot infer that this concrete hungry person should be fed, as she might have undergone a stomach operation and must not eat. Moreover, I might reject my own judgement presented by the intellect on what I should do in these particular circumstances and determine to do something else, or to do even something contrary to my own judgement. This is part of human freedom. I am also able to pass an evaluative judgement on my decision: whether it was right or wrong. I may also regret my decision.

Understood in this way, the decision means self-determination for action – it establishes me as a source of action (an egotic moment) and I recognise this fact (it is my decision). Following Aristotle's expression that "man is a moving principle or begetter of his actions as of children,"[6] we may say that the decision starts a new causal chain in a double sense: (1) it builds me as a certain kind of a person. If I decide to murder someone, this decision starts building me as a murderer. These are inner consequences of any decision one makes. (2) If a decision is executed, it changes the world and "begets" changes in the world. A decision may concern doing something (dancing, climbing a mountain); and it may concern bringing something about – creating a new object or state of affairs that would exist independently of me. The latter case assumes a plan, a vision of a state of affairs seen as desirable. Thus, the decision always brings inner consequences, and when executed, always brings external consequences. In short, by making a decision, I become an efficient cause of changes in myself and the world. The fact that I determine myself to act in a certain way does not mean that I am not influenced by emotion, previous experiences, other people's views, and other factors. Yet taking into account such factors does not mean that they determine me like gravitational force determines my body.

Since it is me who has begotten the change as a parent has begotten children, I am responsible for my decision and its consequences (the ethical moment). This in turn presupposes some knowledge of the workings of the world and therefore both my ability to cognise the world and the intelligibility of the world. The decision – and ascribing responsibility for it – presupposes also freedom to choose among possible actions opened in given circumstances, as well as some external freedom to execute a decision and some influence on the development of action.[7] There are interesting consequences of such a view on decision. For example, the responsibility for a decision involves responsibility for acquiring relevant knowledge, for deliberating rather than making a hasty judgement, having not only knowledge but also understanding. I will not develop this set of issues, as it would drive us away from the main issue of the article. Yet, one more thing needs to be stressed: sometimes we are in dilemma. Dilemmas are situations in which the decision-maker must consider two or more moral values or duties but can only honour one of them. There are many types: conflicting values cannot be realised at the same time, or one value cannot be realised with regard to more than one object. As Aristotle rightly says,[8] we do not need to decide about something that is necessary and/or cannot be changed. Thus, the primary condition is that we recognise that a given case requires a decision. We then need at least: the primary goal we wish to achieve by making a decision; the workings of the world (including people and society); concrete circumstances in which a decision has to be made; alternatives opened, expectations concerning consequences of choosing any of them, and evaluation of those consequences as good/bad, acceptable/unacceptable, desired/harmful, and so on. In short, the decision requires knowledge and deliberation through which reasons for options are weighed. In those

4 See: M. A. Krąpiec, *I—Man: An Outline of Philosophical Anthropology,* trans. by M. Lescoe and others (New Britain, Conn.: Mariel Publications, 1983); M. A. Krąpiec, "Decyzja" [Decision], in *Powszechna Encyklopedia Filozofii* [*The Universal Encyclopedia of Philosophy*], vol 2. (Lublin: Polskie Towarzystwo Tomasza z Akwinu, 2001, 442-44); S. Kamiński, "Wisdom in Science and Philosophy," in *Studies in Logic and Theory of Knowledge*, vol. 1, eds L. Borkowski, S. Kamiński, and A. B. Stępień (Lublin: TN KUL 1985), 91-96.

5 See: M. A. Krąpiec and A. Maryniarczyk, *The Lublin Philosophical School* (Lublin: Polskie Towarzystwo Tomasza z Akwinu, 2010).

6 Aristotle, *Nicomachean Ethics*, trans. by William D. Ross (Oxford: Clarendon Press 1925), Book III.5, 1113b.17.

7 The issue of responsibility and its metaphysical foundations is well disscused in: R. Ingarden, *Über die Verantwortung. Ihre ontischen Fundamente* (Stuttgart: Reclam 1970).

8 See: Aristotle, *Nicomachean Ethics*, Book III.3.

deliberations, both facts and values are intertwined. Moreover, rationality requires that I am able to provide justification for my decision.
Before applying this understanding of the decision to AI, let us first consider what is expected from AI's actions.

The Asilomar Principles

On January 5–8, 2017, the Future of Life Institute organised the Asilomar Conference on Beneficial AI. Its outcome was a set of guidelines for AI research – the so-called *Asilomar AI Principles*.[9] These 23 principles are organised into three groups: research issues, ethics and values, and long-term issues. The starting point for elaborating these principles is the conviction that technology is giving life the potential to flourish, but is also endangering it. We need, then, to develop technology in such a way as to promote the flourishing of the world and prevent disastrous consequences. AI is a powerful technology that has already proved to be beneficial for many people all over the world. The assumption is that "its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead."[10] Some of the principles indicate how any AI should act and only those principles will be here analyzed. The other principles, which regulate the action of researchers, sponsors, and users of AI, are omitted.
Already principle 1 states: "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence."[11] The term *beneficial* here is crucial. The first thing to note is that this principle states that any AI should work in a way that is beneficial. That is, it states that AI itself should *act* beneficially, and not that AI should be *used* by a human agent in a beneficial way. So, this principle assumes that AI systems are autonomous agents, working without a direct command of a human. This is turn presupposes that the term "beneficial" has some objective sense that AI is able to grasp in a given situation and act accordingly. The second thing is that acting beneficially requires more than Isaac Asimov's laws of robotics, formulated in his famous collection of stories *I, Robot*. Asimov's first law states that a robot must not injure a human being or, through inaction, allow a human being to come to harm.[12] Acting beneficially – for the good of a certain being – is more than non-harming, for it requires taking an active attitude to promote the development and well-being of that subject. So, AI should be able to grasp the nature of a given being and a situation it is in, develop possible action paths, and then determine itself to act in a way that serves the good of that being. In other words, AI should work ethically. This fact is widely recognised and this is a reason for the enormous development of AI ethics. Two other principles explain further the meaning of AI's beneficiality. Principle 10 states: "Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation."[13] The principle then suggests that AI's actions should contribute to the realisation of human values. But what are human values? Let us name a few: life, health, solidarity, security, cooperation, peace, friendship, clean environment, biodiversity . . . This list is very short but it nevertheless indicates something important: beings for the good of which AI should act are not just human beings. An AI should protect *me* against a dog attack, but it also should protect *my* dog against the attack of a human; or when shopping, it should choose beverages in glass bottles to save the environment from plastic pollution. Principle 11 specifies further that "AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity."[14] It seems that the term "compatible" is here too weak in comparison with the previous principle, for what is at stake is promoting and protecting human values, and not just some sort of neutrality with regard to them. Yet this principle indicates the nature of the human being: she possesses dignity and rights (presumably stemming from dignity), is free, and freely creates culture (this is why there are many cultures). In short, the two principles quoted make philosophical assumptions. And those assumptions are necessary to establish what actions are obligatory, permissible, or forbidden, if the good of a human person is to be realised. Similar assumptions must be made about other kinds of beings, for – as claimed above – AI should work for the good not only of human beings. Principle 23 contains the most stringent requirement. It states: "Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals and for the benefit of all humanity rather than one state or organisation."[15] The term *all humanity* introduces the issue of future generations, for AI should act *now* to develop and protect the good of *future* persons and more generally, future beings. This, in turn, requires assuming a conception of good human life, for future generations cannot be

9 "Asilomar AI Principles," https://futureoflife.org/ai-principles/ [accessed 16.08.2021].
10 Ibid.
11 Ibid (1).
12 I. Asimov, *I, Robot* (New York: Gnome Press, 1950).
13 Asilomar AI Principles (10).
14 Ibid (11).
15 Ibid (23).

asked what world they would like to live in. The above preliminary remarks on the principles that should govern the AI's action allow drawing a more general conclusion. Marcin Garbowski is right that "The principles presented in the declaration are a good point of departure for further discussion and analysis, but without making the basic ontological criteria more precise and backed by strict regulations these points are not enough to secure a safe relationship between humans and AI."[16] We then need to embed the understanding of AI as acting subjects in philosophy. This is explicitly recognised by thinkers working in the domain of AI ethics: "AI researchers will need to admit their naiveté in the field of ethics and convince philosophers that there is a pressing need for their service; philosophers need to be a bit more pragmatic than many are wont to be and make an effort to sharpen ethical theory in domains where machines will be active."[17] The question immediately arises: since there are many philosophies, which philosophy (not just ethics but also ontology and epistemology) should serve as a basis for AI ethics? AI practitioners use various theories,[18] but in this paper I will not defend a particular answer to that question. Instead, I ask what kind of an agent the AI must be in order to be able to follow the principle specified above.

In Search of an Artificial Wisdom?

Asilomar Principle 16 states: "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."[19] This principle sees AI as "tools" for human goals. But AI systems are quite peculiar tools. Delegation of decisions means that AI systems themselves make decisions without a human command and as "tools" work autonomously. Here there is no difference between a human agent and an AI agent, once decisions concerning certain matters are delegated to them. Following the analysis of the decision developed in the first section, we have to recognise that an AI agent self-determines itself to act in a certain way. As specified in the Asilomar Principles, the AI's decision should be ethically right, i.e., it should realise what is good for a being to which a given action is directed, whatever goal was set. This points to a classical understanding of *praxis*: when one decides what to do, the goal, means, and circumstances of planned action are considered separately. That is, the goal might be good, but all available means are morally unacceptable or the goal and means are morally acceptable but performing that action in certain particular circumstances would bring harmful consequences, and therefore one should refrain from action. The AI goal is specified: serve the realisation of human values. Actions chosen by AI should then be morally acceptable as means to that goal and their consequences should be morally acceptable as well. Thus, AI systems should pass judgements on "this should be done" and then determine to itself "I will do this." Moreover, AI should be able in some sense to justify its decision. This already sets restrictions on how to build "moral machines."

James Moore distinguishes implicitly ethical agents, explicitly ethical agents, and full moral agents.[20] Human persons possess full moral agency. Implicitly ethical agents are "operationally moral," i.e., they do what their designers implemented in them. Explicitly ethical agents are those who make moral judgments that determine their course of action without direct human instruction. In their now classical book *Moral Machines: Teaching Robots Right from Wrong*[21] Wendell Wallach and Colin Allen call that *functional morality*. Most AI ethicists agree that given the scope, unpredictability, and complexity of the context in which AI operates, building AI systems as explicit moral agents is the only reasonable option. And to provide any justification for its decision, AI must appeal to some objective principles: "The ethical component of machines that affect humans' life must be transparent, and principles that seems reasonable to human beings provide that transparency."[22] Yet, if we take into account the understanding of the decision and the content of the Asilomar Principles specifying how the AI should act, an array of problems arises.

Let us start with a thought experiment. Suppose that an AI works as a distributor of respirators. One night there are two patients arriving in the hospital but there is only one respirator left. There are conflicting instructions: "Give it to patient X" and "give it to patient Y." Without a respirator both patients die. What is required is a judgement solving a dilemma of who gets a respirator, and such a judgement should take into account facts and values involved. The AI should then be able to distinguish ethically relevant factors. In the situation described, the values are clear: life and health. But is the fact that one person is a child

16 M. Garbowski, "A Critical Analysis of the Asilomar AI Principles," *Zeszyty Naukowe Politechniki Śląskiej. Seria: organizacja i zarządzanie* 113 (2017): 45-55.
17 M. Anderson and S.L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine*, 28 no. 4 (2007): 15-25.
18 See, for example, L.A. Dennis, M. Slavkovik, "Machines that Know Right and Cannot Do Wrong: The Theory and Practice of Machine Ethics," *IEEE Intelligent Informatics Bulletin* 19 no 1 (2018): 8-11.
19 The Asilomar Principles.
20 J. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21 no 4 (2006):18-21. Available at https://www.researchgate.net/publication/220629129_The_Nature_Importance_and_Difficulty_of_Machine_Ethics [accessed 7.09.2021].
21 W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2008).
22 A. Anderson, S. L. Anderson, op. cit.

and the other an adult ethically relevant? Or that one is a famous writer and the other has Down syndrome? When judgement is passed and followed in action, a human agent might also discover for herself that the judgement was wrong and she may regret it. Who should judge the correctness of a given ethical judgment in the case of an AI agent? The situation becomes even more complicated when we realise that the judgement is individual; that is, we may agree as to the principle of "respect human dignity" and disagree on whether a particular action fulfils that principle. Is euthanasia a sign of respect for human dignity or quite the contrary? It seems that what is required for AI agents that are explicitly ethical is some sort of artificial conscience, not just artificial intelligence with some algorithms for moral reasoning. Otherwise, before buying an AI nurse or babysitter we should ask on what moral basis those algorithms were written to make sure that they agree with ours. Or maybe the usage instructions should include a detailed "moral declaration" of a given AI.

The need for making the moral algorithm explicit is even more vivid when we ask for justification. If the family of a patient who died without a respirator asks why, they should get an explanation of why such a decision was made and who is responsible for it. Can an AI provide such an explanation? To what reason can it appeal? And who is responsible for the judgement of what the AI did? A programmer? The AI itself? The latter answer opens new issues such as punishment and restitution. We need to stress here the difference between ordinary machines and AI. Machines can harm users. In such a case one may inquire whether it was an accident, or a design error, or carelessness on the part of a producer. Yet, this is an AI that makes a judgement of what should be done and follows it. If there are harmful consequences of an AI's action – who is accountable?

The problem is even more complicated when we realise that AI should not be able to do what is wrong. The Asilomar Principles require that AI is a beneficial intelligence, a "good moral agent" that makes the right decisions in various circumstances, correctly solves moral dilemmas, and generally works for the common good. The AI must then be able to recognise facts and values, pronounce for the truth and the good and act accordingly. That is, taking into account morally relevant factors, it should determine what is the right thing to do in these particular circumstances with regard to this particular subject and then follow the judgement. In classical philosophy, this human ability is called *practical wisdom* and is seen as a virtue. It seems then that AI ethics finally aims at artificial wisdom, or at wise artificial intelligence. Some thinkers even claim that AI would be "morally better" than human beings, for human moral judgements are disturbed by emotions, partiality, individual vices. AI is immune to such disturbances. It is also better at collecting data, analyzing facts, discovering alternatives, and so on, so its judgements should be better. Moreover, AI may face moral dilemmas more often; there could be moral dilemmas that humans have never faced because of our limited capacities. So, it seems that human experience would not be sufficient for equipping AI with morality. Should we then study AI's decisions and learn from them?

Conclusions

Against the background of a classical understanding of decisions and the Asilomar Principles for AI development, I showed that in order to have AIs operate as autonomous agents, we need to equip them with some sort of artificial conscience, i.e., the power of recognising ethically relevant facts and values, formulating judgements on what should be done to achieve what is objectively good. The ability to evaluate reality from a factual and ethical point of view, pass morally right judgements, and then willingly follow that judgement is called practical wisdom. I do not debate the technical possibility of creating wise AI systems. Yet, any attempt at doing so should be nested in philosophy, with its ontology, epistemology, and ethics. Otherwise, we may develop AI systems that in certain situations pass contradictory judgements and act accordingly – almost certainly harmful consequences would follow. So, the basic problem is not so much technical as meta-philosophical: which philosophy is the right philosophy for AI that is to operate both in the real world and in virtual realities. The problem is recognised but not debated widely. Developing such a debate seems to be a challenge for both academics and engineers. For only then follow-up problems can be solved. I pointed to some of these problems: discovering ethically relevant factors, justifying decisions, being responsible and accountable for consequences of action. Moreover, once we expect to create wise AI, the question of its status arises. And there is no way to avoid such deep philosophical debates, for AI is developing rapidly. And it seems that to solve those problems researchers, designers, producers and sponsors must themselves employ practical wisdom.

Such a judgement should take into account facts and values involved.

References

Anderson, Michael and Susan Leigh Anderson. “Machine Ethics: Creating an Ethical Intelligent Agent.” *AI Magazine* 28, no. 4 (2007): 15-25. DOI: 10.1609/aimag.v28i4.2065.

Asilomar AI Principles. https://futureoflife.org/ai-principles/ [accessed 16.08.2021].

Asimov, Isaac. *I, Robot.* New York: Gnome, 1950.

Aristotle, *Nicomachean Ethics.* Trans. by William D. Ross. Oxford: Clarendon Press. 1925.

Boston Dynamics. https://www.bostondynamics.com/ [accessed 9.08.2021].

Dennis, Louise A. and Marija Slavkovik. “Machines that Know Right and Cannot Do Wrong: The Theory and Practice of Machine Ethics.” *IEEE Intelligent Informatics Bulletin* 19, no. 1 (2018): 8-11.

Editor. “Russian Employees Fired: IA Labeled Them Unproductive.” https://today.in-24.com/technology/205364.html [accessed 9.08.2021].

Garbowski, Marcin. “A Critical Analysis of the Asilomar AI Principles.” *Zeszyty Naukowe Politechniki Śląskiej. Organizacja i Zarządzanie* 113 (2017): 45-55. DOI: 10.29119/1641-3466.2018.1153.4.

Ingarden, Roman. *Über die Verantwortung. Ihre ontischen Fundamente.* Stuttgart: Reclam, 1970.

Kamiński, Stanisław. “Wisdom in Science and Philosophy.” In *Studies in Logic and Theory of Knowledge*, vol. 1. Ed. Ludwik Borkowski, Stanisław Kamiński, and Antoni B. Stępień. Lublin: TN KUL, 1985, 91-96.

Krąpiec, Mieczysław A. *I—Man: An Outline of Philosophical Anthropology.* Trans. by M. Lescoe and others. New Britain, Conn.: Mariel Publications, 1983.

Krąpiec, Mieczysław A. “Decyzja.” In *Powszechna Encyklopedia Filozofii,* vol 2. Lublin: Polskie Towarzystwo Tomasza z Akwinu, 2001, 442-444.

Krąpiec, Mieczysław A. and Andrzej Maryniarczyk. *The Lublin Philosophical School.* Lublin: Polskie Towarzystwo Tomasza z Akwinu, 2010.

Mizokami, Kyle. “For the First Time, Drones Autonomously Attacked Humans. This Is a Turning Point.” https://www.popularmechanics.com/military/weapons/a36559508/drones-autonomously-attacked-humans-libya-united-nations-report/ [accessed 9.08.2021].

Moor, James. “The Nature, Importance, and Difficulty of Machine Ethics.” *IEEE Intelligent Systems,* 21 no. 4 (2006): 18-21. DOI: 10.1109/MIS.2006.80.

Wallach, Wendell and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong* Oxford: Oxford University Press, 2008. DOI: 10.1093/acprof:oso/9780195374049.001.0001.