Global Journal of Arts Humanity and Social Sciences ISSN: 2583-2034

GSAR PUBLISHERS
GLOBAL SCIENTIFIC AND ACADEMIC RESEARCH

Glob. J.Arts.Humanit.Soc.Sci
ISSN: 2583-2034
Vol-2 Iss-11, page 828-832

IN WHAT SENSE ARTIFICIAL AGENTS SHOULD BE A SUBJECT TO MORAL JUDGMENTS?

BY

Iaroslav Petik

Senior researcher Museum for Outstanding Figures in Ukrainian Culture Ukraine

Article History
Received: 12/11/2022
Accepted: 16/11/2022
Published: 18/11/2022

Corresponding author:

Iaroslav Petik

Abstract

This paper asks the question whether artificial agents are subject to moral judgments. In other words can robots be assessed as good or bad and does harming them bring moral and legal consequences. Moral judgents is a good term which also entails such notions as “rights”, “dignity” and personality. Most of the essay deals with the possible criticism of the initial statement which falls into two distinct categories.

The first criticism denies the personality in artificial agents. Author points out both to the general case of such a denial and to particular cases, giving counter arguments to each one. The second criticism concerns empathy and compassion which are the foundation for legal and moral norms regulating communication between the individuals. The idea according to which robots are equal to humans and should not be treated as pets is defended.

The main conclusion of the paper is that artificial agents should be the subject to moral judgeents and should be treated equally to humans. That is both a moral and political position.

Key words: artificial intelligence, robot, ethics, morality, applied ethics;

INTRODUCTION

Artificial intelligence and a variety of ethical problems surrounding it is at the scope of attention of philosophy for many years already. Are robots personalities? Do they have equal rights with humans? Does an artificial agent have a soul?

All these problems are not just in the domain of moral philosophy but also in the domain of philosophy of mind. It is due to the last we define a personality, consciousness, and mind and artificial analogies of these concepts (if there are).

Pop culture also makes a lot of references to this particular topic. Robots and conflicts in society caused by artificial intelligence is a main topic for numerous sci-fi books, films, and tv-shows. Not all of them are helpful when it comes to philosophical and moral analysis of the issue but some of them bear valuable insights. Among the later ones “Blade Runner” and “Westworld” can be named. However, this essay will remain in the scope of academic philosophy taking only some ideas from pop-cultural context and not trying to analyze that context as it is.

In order to keep the question solely inside the borders of moral philosophy some prior definitions should be made. The initial question in the headline “In what sense artificial agents should be a subject to moral judgments?” is the product of few of those narrowing definitions.

Why “artificial agent” and not “robot” or other synonyms? Term “artificial agent” is more neutral and describes the wider scope of different entities. Robot is an artificial agent as well as the sentient software program (artificial intelligence). “Artificial agent” also applies to the entities created with the help of biotechnologies like replicants from a famous film “Blade Runner”. Despite being neutral this term also does not attack or humiliate entities it refers to which is logical as the author leans towards admitting rights and dignity of artificial agents.

Why question “moral judgments” rather than “soul” or “dignity” or “rights”? The question about “soul” sounds most interesting but it brings a lot of problematic context with it. Does the “soul” exist? If yes, what is the “soul” of a being? What if personally I (no matter

if wrong or right) do not believe in the existence of the "soul"? These are very big questions thinkers and philosophers tried to answer throughout the history of humankind. If we enter the domain of these problems it is doubtful time for considering the "artificial agent" side of the question will ever come. Besides, it refers more to theology than philosophy.

Terms like "dignity" or "rights" pick up only some sides of the problem. Does having dignity entail having rights? Or vice versa? It also is extremely background-dependent as different law systems define rights and personal dignity in different ways.

On the other hand "being a subject to moral judgment" explains ideally the core of the deal. If the entity is a subject to moral judgment you consider it equal, having dignity, rights, and so on. Or, at least, if not so, you can explain in what particular sense this entity is a subject to moral judgments. If you consider an entity a moral agent this brings most of the other consequences such as legal and some social conventions. Therefore, the term "moral judgement" is the best representation of the issue.

In my opinion, artificial agents do have their rights and dignity, the same as humans and therefore they are subjects to the same moral judgements as regular humans are. The two views which oppose this opinion of mine are that 1) artificial agents are not subjects to moral judgements2) artificial agents are subjects to moral judgement in another sense than humans.

These views are interconnected and I will examine both of them in detail in order to give clear and logical arguments against them.

Artificial agents are not subjects to moral judgements at all

Defenders of this radical view claim that machines or other artificial agents do not possess defining quality which makes them equal to human personalities. Depending on the background of the particular defender this defining quality may be mentioned "soul", "dignity", "personality" or any other personal trait connected to the values of that particular defender. It can be generalized that all of these traits in some sense make an artificial agent a personality, an entity that should be considered as an equal to a human being. These traits may be called metaphysical "personality-making" traits.

However, it should be mentioned that there is at least a logical possibility that there exists such a radical defender of this view who despite recognizing personality amongst artificial agents still considers them to be treated as non-humans. This seems like the ideology of a modern slavery proponent. That is a cynical type of fascism that is senseless to oppose, at least within philosophical discourse.

On the first sight, the opposing strategy against each of the mentioned special "personality-making" traits should be different which brings huge difficulties for such an opposing. You need one tactic to prove that artificial agents have "souls" and completely different tactics if the issue in question is "dignity". However, as was already mentioned this "system of traits" resembles the metaphysical hierarchy. Moreover, in a sense, it is such a metaphysical hierarchy. The tactic should be in accordance with that fact.

There is an agent which should be considered a personality (given particular ontological status) if there is a special "personality-maker" in existence for him. Therefore, the strategy for opposing "anti-robot" views is next. First to attack the general "personality-making" metaphysical hierarchy and secondly to consider the consequences for the particular cases.

We start from comparing a human and an artificial agent. An artificial agent as described by futurists can engage in any type of human activity as well as human and sometimes do it even better. This means that at least in his intellectual capacities the artificial agent is similar to a human. It is practical to view intellectual activity first of all. Most of the physical activities are easily simulated by simple mechanisms while intellectual activity is strong evidence for the serious level of capacity of the artificial agent.

The "anti-robot" ideology tactics is to claim that any activity of the artificial agent is only copying the human activity including intellectual activities. In other words, "robot" does not have the intellect the same as a human personality, it only has the copying mechanism or something else.

The problem of defining intellectual activity and copying of the intellectual activity is quite an old problem in philosophy of artificial intelligence. The famous Turing's "Imitation Game" conception shows that any criterion for the border between those two phenomena is arbitrary (Petzold, 2008). The scenario proposes a situation when a human expert communicates in one session with both the machine and another real human person. If an expert cannot differ between the two then the machine participating in the experiment is sentient. Another famous conception is the "Chinese Room" thought experiment which also attacks that criterion (Searle, 1984).

"Chinese Room" proposes a translation scenario which precisely follows syntactical rules but completely ignores the semantics. This thought experiment demonstrates the absurdity of the very idea of the copying mechanism for the human intellect for reaching sentience. However absurd that is, contemporary artificial intelligence technologies are based precisely on the idea of plain copying the operations of the human intellect.

Nevertheless, this experiment speaks in favor of those opposing the "criterion principle". In other words if "Chinese Room" is acceptable so is the idea that there is no difference between the copying mechanism and human intellect. If there was such a difference - it would be easy to differ a machine from a man and "Chinese Room" would have been a weak argument.

The copying mechanism criticism resembles the metaphysical conception of "personality-making" trait. In this case, an intellect is that trait. The criticism then has two distinct parts. Firstly, it claims that there is no such trait at all. Robots do not have an intellect. Secondly, it may claim that despite effectively engaging in intellectual activities of human personalities it does not have the same intellectual mechanism but a completely different one. This different mechanism on its part cannot be considered a proper "personality-making" trait.

The first part of the criticism is answered straightforwardly. Artificial agents are producing the same activity as the human personality. As we abstract from all other traits of a human personality and center only on his activities there is no other way to distinguish personality and not a personality. Our initial assumption is that an artificial agent engages in all activities of the human with the same level of effectiveness or even above. Therefore, there is no other logical variant except for admitting artificial agents being the same personality as a regular human.

For the sake of experiment, let us assume that in the situation of considering activity as the only parameter there is some hidden difference between two agents engaged in that activity. In this situation, we have to admit that there is no way to differentiate between regular humans besides "human and robot" cases. Therefore, this imaginary situation is disproved by reductio ad absurdum.

The second part of the argument is much more sophisticated. An agent engages into an activity and provides the same or a better result but does so due to the completely different inner mechanism. In this sense, he is similar to the factory machine. Factory machine does the same job as a single human worker, often more effectively. However, it is a machine, it is evident it shouldn't be given the same rights as the human worker. It is the same if we gave the separate mind and emotional world to the primitive tools our ancestors used. A rock or a stick do not have the personality. There is a radical view of panpsychism but it faces too many problems and it should not be dealt with in this essay.

The case with the intellectual activity seems to be described by this metaphor as well. Nevertheless, is it really so? The factory machine should not be given rights and admitted to have dignity because it does not feel or think. It is part of the neutral material nature. Therefore, the problem brings on a completely different context. That is empathy and compassion we have to other human beings due to them possessing the same natural psychological traits as we do. These traits are thinking, self-awareness, feelings, and emotions.

However, in the previous paragraphs, the intellectual activity of an agent was analyzed in particular. Intellectual activities presume some inner mechanisms for an artificial agent, at least much more complicated than for a simple factory machine.

A critic might say that a complexity of an intellectual mechanism does not matter in the case of empathy. There are software programs that make enormously complex calculations sometimes impossible to do for a simple human mind but no one ever calls those software programs sentient or possessing feelings.

In this particular place, philosophy of artificial intelligence is tightly connected to philosophy of mind. It is philosophy of mind, which defines what is consciousness, mind, and freedom of will. Nevertheless, the problem is that there are many competing conceptions about consciousness or similar notions, which often contradict one another. If we adopt one of those conceptions how can we justify adopting that particular conception? Moreover, how should we defend it against existing criticism?

Let us assume we have adopted the conception of consciousness by philosopher David Chalmers (Chalmers, 1997). It is quite good for our purposes because it provides a distinct criterion for distinguishing sentient and non-sentient beings. Chalmers also wrote about distinguishing the things of the material nature and sentient beings.

Therefore, after adopting this ideology we should provide an answer to main criticism. For example, we should defend consciousness as the criterion. This essay then will be more about philosophy of mind than philosophy of artificial intelligence.

Different conceptions on their part are not worse than Chalmer's philosophy. Daniel Dennett, for example, adopts a very interesting view according to which there is no "personality-making" trait at all (Dennett, 1982). Dennett claims that it is a question of a specific stance. You adopt a special stance towards an object in order to consider it a personality. You do it for the reasons you consider rational. Therefore, if you play chess with a computer program you can consider this program sentient as it is just a question of a mere stance.

That resonates with a judgement of one of the characters in "Westworld" tv-show. I cite this show because despite the pop-cultural nature it is quite deep and intellectual bringing a lot of valuable judgement about philosophy of artificial intelligence. Professor Ford says that there is no defined border crossing which an artificial agent becomes human. In fact, he sounds very ironic about people who believe that such a border exists. That is very similar to Dennett's notion of stance.

Therefore, there are two polar opinions about the intellectual criterion as the "personality-maker". First, there is such an intellectual mechanism that makes a personality, and the second as it was told, "there is no such border". If it is later, there is no sense in arguing – with certain modifications, even the mentioned factory machine can be finally considered sentient. However bizarre this view may seem, it exists.

Philosophy of mind is useful as the source of conceptions for the philosophy of artificial intelligence and not as a solid doctrine that should be defended or negated prior to any ideas concerning the status of artificial agents.

So while considering the first of mentioned possibilities it is clearly seen that it is about that particular intellectual mechanism. There is a human personality who engages into certain intellectual activity due to his intellectual mechanism and his intellectual mechanism is a "personality-making" trait. Then there is a machine, which also engages into that type of activity but due to another intellectual mechanism, which cannot be considered such a trait. We take only intellectual activity into consideration due to the fact that most of the physical activities are already effectively simulated by non-sentient machines.

In our analysis, we came to the same situation that made Turing design his "Imitation Game". In short, if there is no vivid difference between the intellectual capacity of the artificial agent and a human it should be concluded that an artificial agent is as

sentient as that human. There are of course also numerous attacks on the concept of a border including "Chinese Room".

Chinese Room underlines the concept of copying the intellectual activity making it look completely absurd. Nevertheless, isn't it already proved that a concept of a defined criterion is as absurd?

Therefore, it can be concluded that the very concept of the "different" intellectual mechanism is flawed. It is senseless to theorize about it. If a machine engages into intellectual activity and there is no way to tell it apart from a regular human and this activity cannot be simulated then this artificial agent should be admitted to be sentient.

Turing's example seems to be the most rational criterion. If a human person fails to distinguish an artificial agent based on his intellectual activity then it should be considered to be sentient. The only "hard" part is the expertise of the human person taking part in the experiment. He or she should be an expert but how to define that expertise is a question of practice, not philosophy.

Sentiency on its part as such a "personality-making" trait entails artificial agents to be subject to moral judgements the same as the regular human person. This means that he can be morally evaluated as being good or bad. Besides, harm made to him brings the same moral and law consequences as harm to a regular human.

The evaluation of the "personality-making" traits metaphysical system is finished. It is time to assess other variants of personal traits in the light of the abovementioned arguments.

Most of them have a common root in the empathy of human beings. We prescribe "rights" and "dignity" to other human beings because we believe they can feel the same as we do. As was mentioned in the beginning of the paper that establishes a strong connection of these notions with values of a particular agent.

However, as was shown criticism built on "metaphysical traits" is contradictory and incoherent. Therefore the "empathists" should relate only to the psychological side of the problem. "Soul" or "dignity" is based on compassion which on its part has degrees. This is exactly what should be analyzed in the next chapter of the essay concerning the second possible opinion about artificial agents and moral judgements.

Artificial agents are subjects to moral judgements in different sense than humans

What is primarily meant by "different sense" of moral judgements? Previous chapter also has spoken about the degrees of empathy. In this case we presume an artificial agent is a subject to moral judgements. Therefore he is sentient and can be considered good or bad, harming him may have consequences et cetera. However, the moral judgements are different from those concerning humans. The closest explanatory metaphor is the case with animals.

Most of the progressive society has compassion towards animals and admits that irrational harm caused to them should have consequences. There is much empathy to it. However, there are numerous cases of harming animals due to rational reasons (for example killing animals for meat and skin) and irrational (purely sadistic) reasons.

Killing an animal for irrational reasons is condemned as a crime and a moral misdoing but not the same as killing a human. In this sense, we speak about the degree of moral judgement. We have compassion for animals but it is not the same as compassion towards people. Human persons are the same as we - they think, they feel, they have dignity. Animals also feel (it is a question whether they think the same way we do) but they are more primitive than we are. Therefore our compassion for them goes to a lesser degree. So are the moral consequences.

Perhaps a more developed society of the future will have a different opinion on this matter but today's situation is outlined. The question is whether this approach is valid for artificial agents? Previous chapter provided counter-arguments to "anti-robot" views and protected the position according to which artificial agents are subjects to empathy and compassion. Nevertheless what if robots like animals have that compassion to a lesser degree?

This view is much more sophisticated than simple denying of the status of the artificial agent. After all artificial agents are created by us, people. Therefore a somewhat theological problem arises.

In the previous text, it was assumed that artificial agents have intellect of the human level or above and that there is no "hidden criterion" to distinguish human intellect and that of the artificial agent. At least in this way machines are not inferior to men and often even inferior.

Again turning to the pop-cultural context: "Blade Runner" explicitly tells that replicants often surpass humans both in their professional skills and depth of emotional world. Nevertheless, they are creations of humans and therefore their purpose is defined by humans. Their memories and personality are artificial.

Why theology? The concept of connection of creator and a creation goes to the monotheistic religions and theological problems surrounding it. God created people and he is also a source of the purpose of our existence. Whether it is a divine mission or a simple moral code, the creator defines the aim of the life of his creations.

Philosophical metaphysics is also acquainted with that kind of idea. Aristotle wrote about the final purpose of things which lie in them potentially (Aristotle, 2016). However, there were those in metaphysics who criticized Aristotle. Maybe there is no potentiality in things. At least that is a debated metaphysical topic and there are different opinions on that.

Purely theological context of a problem may be addressed in two different ways. First of all the mentioned creator-creation concept refers only to a few monotheistic religions. Pagan ancient Greek religion sees it in different perspectives, for example. People are indeed creation of gods and own them a deal but creation in general, as well as gods in particular, comes from the Chaos. Therefore the higher purpose of life is still unknown.

Secondly, the notion of freedom of will which is extremely important for monotheistic religions comes into conflict with the predestined purpose. If God gave us the ability to choose, does this mean there are no predestined events and there is no purpose in our existence? Even if there is a moral code existing there is always a

possibility to violate that code. The only option for an "anti-robot" thinker here to defend is to say that there are only two possible paths for creation to follow: the righteous and unrighteous one.

That is an interesting theory which applied to our situation results in admitting there are artificial agents who are sinners. However funny and fresh this may seem it is self-contradictory. If robots are sinners they have their moral code and their purpose which they acquired from a higher being than a simple human, even if he is an extremely talented engineer. However "anti-robot" you are, you do not consider computer engineers the right entities to provide artificial agents with their moral code and place in the Universe.

Then there is Aristotle's metaphysical conception that left. It is not "anti-robot" as the previous points of view. Anything can bear some potentiality in it. Including the things of nature that were not created by humans. However, if we follow "anti-robot" ideology, artificial agents are particular human-created things. To counter that it is enough to ask some of the initial questions about Aristotle's metaphysics. According to this philosopher, all the things acquire their movement from the First Engine. It is logical to assume that the purpose of existence is also distributed among the things according to some great predestined plan including artificial agents. Therefore the purpose for robots is given by someone or something ontologically higher than humans.

The main evidence for not considering artificial agents inferior towards people due to the history of their creation is outlined. All other types of arguments concern the criterion of an agent being sentient which on its part was fully covered in the previous chapter.

However, there is a type of point of view according to which artificial agents are subject to moral judgements to a higher degree than humans. It may seem to sound odd but it is at least conceivable.

Everyone has a right for their own beliefs of course as soon as they are not harming others. It is possible to value some robots more than humans especially if those robots are smart as humans. However, the principle of justice requires all beings having dignity to be treated equally well.

CONCLUSION

Artificial agents are subject to moral judgements to the same degree as humans are. They can be named good or bad and harming them will bring both legal and moral consequences. The reason for that is they being as sentient as humans and deserving the same degree of empathy.

On the way to admitting this, there are two main groups of "anti-robot" critical arguments. The first one casts doubt on the very core of sentience and intellect of artificial agents. The second mostly concerns the empathy and compassion towards other beings on which moral and legal norms are based.

Both these groups of counter-arguments have their flaws and both are criticized in the two previous chapters. In general, the essay brings evidence for the position according to which truly sentient artificial agents have their dignity and are equal to humans both in moral and legal senses. More to that, the author believes that this position is the only position to which truly enlightened society of the future should stick to.

REFERENCES

1. Aristotle. (2016). *Metaphysics.* Hackett Publishing Company, 712 p.
2. Chalmers D., (1997). *Conscious Mind: In search for fundamental theory.* Oxford University Press, 432 p.
3. Dennett D., (1982). *The Intentional Stance.* The MIT Press, 400 p.
4. Petzold C., (2008). *The Annotated Turing: A Guided Tour Through Alan Turing's Historic Paper on Computability and the Turing Machine.* Wiley, 384 p.
5. Searle J., (1984). *Minds, Brains, and Science.* Harvard University Press, 107 p.