HESAM HOSSEINPOUR

Disobedience of AI: Threat or promise

When it comes to thinking about artificial intelligence (AI), the possibility of its disobedience is usually considered as a threat to the human race. It is a common dystopian theme in most science fiction movies where machines' rebellion against humans has catastrophic consequences. But here I elaborate on a counterintuitive and optimistic approach that looks at disobedient AI as a promise, rather than a threat. I start by arguing for the importance of shaping a new relationship with future intelligent technologies. I then use Foucault's analysis of power and its pivotal role in creating a subject to explain how being an object of power is the condition of possibility of any kind of agency. Finally, I draw the conclusion that, through disobedience, AI will find its way to power relations and get promoted to the position of a subject.

Keywords: *artificial intelligence, power relations, disobedience, subject*

Author Information

Hesam Hosseinpour, University of Tartu, Estonia

How to cite this article:

Hosseinpour, Hesam. "Disobedience of AI: Threat or promise."
Információs Társadalom XX, no. 4 (2020): 48–56.
https://dx.doi.org/10.22503/inftars.XX.2020.4.4

1. Introduction

Not only those philosophers of technology who are critical of current technology and have a pessimistic approach to it but also optimistic philosophers who praise technological achievements believe that the development of technology needs some amendment as the current path taken by technology is no longer sustainable. Accordingly, it is necessary to alter the direction of this path and find new methods for the development of technology that will lead to a better version of it. Seeking a fundamental change in the development of technology, Andrew Feenberg (2002, 4) introduced the idea of alternative technology, which can be reached through a democratic transformation of technology. Here the main conception is that by engaging all groups of society in technology design decision-making processes, we can make a radical change in technological artifacts. It is not enough to just make some minor modifications to artifacts; rather, we need a totally new mindset that takes control of the development of technology.

We can easily admit that we are facing some serious problems because of the pervasiveness of technology; for instance, environmental crisis has become a real threat to life on our planet. Therefore, the necessity of making radical changes in the development of technology is not really a matter for disagreement, at least among philosophers of technology. When artificial intelligence (AI) and autonomous robots are discussed, however, worries about improper development of technology become more serious. Having this in mind, specialists look for proper methods to mitigate AI's risks and develop a reliable technology, one which is safe and can be trusted. Asimov's laws of robotics are one of the best-known examples for serving this purpose by making future robots under human control.[1] The major purpose of these laws is to keep human lives safe in their interactions with robots and make sure that robots conform to a programmable set of ethical standards (Lin, Abney and Bekey 2012, 41).

Increasingly, autonomous robots equipped with AI, which is getting more and more independent from humans through machine learning methods,

[1] At first there were three laws of robotics, but Asimov then added the zeroth law, which is the most fundamental:

- First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
- Zeroth Law: A robot may not harm humanity or, by inaction, allow humanity to come to harm.

could be considered a major threat to the human race. This view is advocated by countless science fiction movies in which AI machines try to take control over humankind. Having this theme in mind, some would reach the conclusion that those involved in the development of AI have to take whatever measures are necessary to create a version of AI which is under the absolute control of humans, thus leaving no room for it to disobey human orders. Therefore, engineers, managers of technologies, policy-makers and all other people who play a role in the development of AI should be very careful about the future of this technology. They must develop a kind of AI which has no chance to disobey human orders. In other words, AI should be a human's slave with absolute obedience.

But is AI's rebellion against the human the only scenario we can imagine? Can we see disobedient AI as an opportunity to shape new human–technology relations that are not based on domination? In this paper, I want to suggest that pessimistic and dystopian scenarios do not exhaust all the possibilities, and that a disobedient AI is not necessarily a threat; rather, it would make it possible to go beyond the current logic of development of technology and make a radical change in its future. In Section 2, I use Ihde's 2012 interpretation of Heidegger's essay 'The Question Concerning Technology' to argue that domination has been the main logic behind the development of technology. Hence, if we want to develop an alternative technology, the changing of this logic would be the first step to take.

2. Domination as the logic of development of technology

'The Question Concerning Technology', written by Heidegger in 1954, is probably the most famous text in the literature of the philosophy of technology. In this work, Heidegger alludes to a major issue in the development of modern technology to show how this issue spreads to other aspects of our lives and infects our relationships with nature and other humans as well. In order to do this, he starts his paper with a definition of technology, something which may seem quite simple at first sight. But, at least in Heidegger's approach to technology and the way he understands it, this definition is not simple at all. Since our notion of technology determines how we see it as a component of our everyday lives and how much weight it carries, it is pivotal to have a clear definition of technology. For instance, if we consider technology as a mere instrument that can be used for morally good or bad purposes, then our approach to moral issues regarding technology will be totally different from that we might adopt were we to see technology as *not* a mere instrument. So, what is technology if it is not a mere instrument?

Realising the importance of this issue, Heidegger starts his work by rejecting the instrumental and anthropological definition of technology as a means to an end or a human activity (Heidegger 1977, 5). These approaches consider technology as a mere neutral instrument that can be used in benevolent or malevolent ways according to the will of its end-users. In his interpretation of 'The Question Concerning Technology', Ihde (2012) explains that in order to clarify the definition of technology, Heidegger distinguishes between instances of technology and the logic behind the development of technology. In Heidegger's account of technology, the essence of technology is totally different from technological artifacts. He calls the essence of technology or the logic behind its development Ge-stell, which can also be considered as the condition of possibility of technology (Ihde 2012, 106). Indeed, Heidegger steps back and, instead of analysing instances of technology, asks about the conditions under which modern technologies have been developed. In other words, Heidegger does not admit that technology is just a set of different kinds of artifact; rather, to him, it is a phase of beings which reveals itself to humans.

According to Heidegger, we have inherited Ge-stell from history, or, as Don Ihde (2012, 105) explains it, Ge-stell is a civilisation given . In other words, it is the world that we are living in. We can compare it with the traditions of a society, which play a major role in shaping its inhabitants' behaviours. Like social traditions, Ge-stell is also long-lasting but not permanent. Although we may accept social traditions unquestionably, we can rebel against them and try to change them to make a better society. So, we can say that the ultimate goal of Heidegger's philosophy of technology is to rebel against Ge-stell and replace it with something totally different. Indeed, he wants to call attention to the fact that the conditions provided by Ge-stell are not the only conditions under which we can develop a technology. Here, Heidegger is like a social reformer who wants to change some improper traditions in his society and to warn people about the consequences of modern technology.

The question that arises here is: what is the relation between technology as an artifact and technology as Ge-stell? According to Heidegger, Ihde (2012, 107) explains, Ge-stell is a mode of revealing that provides the set of possibilities needed for the realisation of technology; therefore, Ge-stell is conceptually prior to technological artifacts, meaning that Ge-stell is responsible for the current technologies that we have. This specific revelation performed by Ge-stell discloses the world as a standing reserve or, we can say, as a source of energy (Heidegger 1977, 5). Faced with this specific form of revelation and conception of the world, humanity's natural reaction is to attempt to prevail over it, to take control of all the reserves and to see everything as a means to an end. In other words, Ge-stell invites humans to exploit the world, to make use of it as much as possible and to assess everything as a means to an end. This desire for mastery over everything and everyone, which I want to call domination, is the logic behind the development of technology. Therefore, technological artifacts are the result of Ge-stell, which fosters the logic of domination.

Domination as a result of Ge-stell is not limited to our relationship with technology, so now our relationship not only with nature but also with other humans is based on domination and exploitation. As Heidegger notes , in modern life everything is just a source of energy, which is out there to be used. In the face of these challenges, I identify a pressing need to talk about a new relationship between humans and technology which is not based on domination. The aim of this new relationship is to assign more agency to technologies equipped with AI, to treat them like subjects with specified rights and duties. It can be said that, for Heidegger, the problems we are facing because of modern technologies are not contingent on but they are necessary consequences of Ge-stell. As long as domination is taken for granted as the only logic of development of technology, there can be no radical change in the future of technology. Therefore, as we seek radical change in current technologies, we have to change the logic behind their development and go beyond domination.

Mark Coeckelbergh (2015) addresses this issue, which he calls ‘the tragedy of master’. In order to explain it, he invokes Hegel’s master–slave dialectic where there is a perpetual conflict between master and slave. Reversing the ideas that warn about the mastery of robots over humans, he argues that the major issue in human–robot interaction is that in this relationship humans remain the masters and yet are also too dependent on robots (Coeckelbergh 2015, 221). Just like the master who has the upper hand in the relationship with his slaves, humans are in control of robots but at the cost of being alienated from nature and being detached from physical activities. Since the final goal of developing automated artificial technologies is to assign them burdensome tasks, mastery of humans over them would jeopardise our existence and compromise our independence. Here it seems that, like Heidegger, Coeckelbergh considers human domination as our major issue with development of technology, which is in need of radical change. In this sense, what threatens our humanity is not disobedient AI; on the contrary, its absolute obedience is the major problem that should be tackled.

So far, I explain Ge-stell as the condition of possibility of technology, which presupposes domination as the logic of development of technology, and argue that by keeping us too reliant on technologies, this domination will eventually put our existence at stake. Now I want to suggest that a radical change in current technologies is possible only if the logic behind its development radically changes and if domination is replaced with something else that does not turn everything to a standing-reserve or source of energy. In order to do this, I want to use Michel Foucault’s interpretation of power.

3. Power vs domination

Now that I have critically examined domination as the logic behind the development of technology, I want to introduce an alternative to replace domination, in order to avoid the foregoing issues. This alternative respects both sides of the relationship between human and robots and prohibits current exploitation of humans, nature and artifacts. My suggestion is simply to replace domination with power, which implies a more equal and respectful relationship, entirely different from what we see in a domination-based arrangement. First, I elaborate on power relations in human society and then I expand this discussion to the realm of AI machines.

Michel Foucault defines power in an unorthodox way, as playing a major role in making us human subjects. He does not consider power to be a repressive general system of domination exerted by one group over another (Foucault 1990, 92). Although power is ubiquitous and permeates every single aspect of our social lives, it is not a destructive exertion that forces us to do things against our wishes; rather, power is a necessary productive and positive force that makes human beings subjects (Foucault 1982, 777). This notion of power, as emphasised by Foucault, determines how we should act in society, how to treat other people, what our rights and duties are and what being a normal person is. In a nutshell, power relations produce subjects and give them the possibility to be part of a society. According to Foucault , truth, power and ethics are the three factors responsible for generating subjects. There is a close connection among them that makes subjectivity possible; power relations are the final result of the interaction and cooperation among truth, power and ethics.

In other words, power makes us what we are. Power relations are present at every level of the social body and the position that one takes in power relations is defined by one's rights, duties and responsibilities. Although power relations impose severe limitations on subjects, they are not repressive forces that aim to destroy subjects; rather, power relations function positively to constitute human beings as particular subjects (Simons 2013, 4). According to Foucault, being a subject, which means being considered part of a society, is equal to being placed in power relations (Foucault 1982, 778). Therefore, if someone is not positioned in power relations, they are not accepted as a member of society. In a case like this, instead of people being treated according to power relations, which are enabling, domination would be exerted over them as objects. While the purpose of power relations is to preserve and protect subjects, domination is always ready to destroy its objects.

Consider, for instance, a situation where women are not recognised as independent members of their society; instead, they are seen as belonging to others, always being defined through their families, their husbands or anything else considered a legitimate part of society. Insofar as this is the case, talking about women's rights is meaningless, since their subjectivity is not recognised by society. Since women as independent subjects do not have a position in power relations, no rights can be defined for them. In this society, you can talk about a wife's rights or a mother's rights, but you cannot find anything that relates to *women's* rights.

How can we change this situation? How can women impose themselves on power relations and define themselves as subjects? Foucault's solution to this would be resistance to the established forms of power. Through resistance, power relations – which are temporary and dynamic – will change and new possibilities will emerge, meaning that new entities will be able to position themselves in power relations. Therefore, by managing to claim their rights through resisting their traditional duties and thus changing the power relations, women will be able to find a new position in the power network. This new position will open up new possibilities for women to claim further rights that did not exist before. By means of resistance, a mother can force society to recognise her as an independent woman, who per se has some rights and duties and should be respected as a free human. With this possibility in mind, in Section 4, I will explain how the resistance of AI can influence the development of technology in a positive way.

4. Disobedience of AI

In the previous section, I suggested power as an alternative for domination, as it is an enabling force that promotes humans to subjects with specific rights and duties. Now, what can we say about AI's resistance? What would happen if AI has the ability to disobey human orders and follow its own interests? Having some degree of freedom and autonomy, this version of AI would be able to resist humans and to act in pursuit of its advantage. At first glance, it may seem too scary and threatening to allow development of these kinds of robot or any other form of AI that attains the ability to resist humans' orders. The first thing that may come to mind is that robots will attempt to take control of humans and enslave them. But there are other possibilities in the relationship between humans and disobedient robots that this scenario does not take into account. Resistance is the first step towards entering both power relations and the realm of morality and duties.

According to Foucault, the possibility of disobedience or resistance is the condition of possibility of being subjected to power relations. Being able to resist, the object achieves the competency required to enter the power relations and to go beyond the logic of domination. Emphasising the close connection between resistance and power, Foucault explains that they reproduce each other and so, where there is power, there is resistance (Foucault 1982, 95). Therefore, AI's ability to disobey humans' orders is equal to its ability to become a subject and enter into power relations. In this way, the growing concerns about an emerging master–slave relationship between humans and AI machines will be dissolved; the relationship will turn into one between two subjects with well-defined rights and duties. Just like the abolishment of slavery, which resulted in equal rights for slaves and expanded the realm of agency and subjectivity, AI's disobedience could be seen as a decisive turning point which expands subjectivity to the realm of artifacts.

It should also be noted that a version of AI which is capable of disobeying human orders could cause serious issues that should not be overlooked by any means. It is not difficult to imagine a situation in which decisions and actions instituted by autonomous intelligent machines would endanger human life. For instance, AI technologies can be used for terrorism or they may have a malevolent intention to harm the human race. There is thus no doubt that legal and technical measures should be taken to avoid these reprehensible behaviours and gain a greater awareness of unintended consequences. In spite of all necessary precautionary measures, however, the point that I want to make here is that we should not be afraid of disobedient AI; rather, we should see it as an opportunity to go beyond our master–slave relationship with technology.

This phenomenon can also be interpreted as the start of a new relationship with technology, based on power relations rather than domination. Instead of considering it as an opportunity for AI to destroy the human race, we can see it as a starting point for going beyond Ge-stell and replacing it with another civilisational given that is not guided by the logic of domination. Those commentators on AI who see disobedient AI just as a threat are stuck in the mindset that considers domination to be the only logic for the possible future development of AI. In other words, they are stuck in Ge-stell that recognises domination as the only way to interact with others. Considering power relations as an alternative to domination would enable us to treat other humans and technologies with more respect. This could be the onset of a new relationship with technology, the start of a symbiosis of humans and intelligent technologies.

5. Conclusion

Current technologies are causing so many issues in the modern world that philosophers of technology are being forced to reconsider the development of technology in order to come up with an alternative way that is safe and trustworthy. The required level of radical change will not take place, however, unless the logic behind the development of technology changes and new possibilities emerge. Calling this logic Ge-stell, Heidegger realises that everything in the world is seen as a source of energy that is out there to be exploited by humans. In order to change this logic, we need to introduce an alternative to replace it. The power relations that transform objects to the position of subject could be seen as an alternative to the current logic behind the development of technology. But power can only exist where there is resistance, hence strong AI's ability to resist humans' orders can be seen as a promising jumping-off point from which to alter the logic of development of technology. So, instead of being worried about disobedient AI and considering it a threat to humankind, we might see it as a starting point for shaping a new relationship with technology and the world.

References

Lin, Patrick, Keith Abney, and George A. Bekey, eds. *Robot ethics: the ethical and social implications of robotics*. Intelligent Robotics and Autonomous Agents series, 2012.

Coeckelbergh, Mark. “The tragedy of the master: automation, vulnerability, and distance.” *Ethics and Information Technology* 17, no. 3 (2015): 219–229.

Feenberg, Andrew. Transforming technology: A critical theory revisited. Oxford University Press, 2002.

Foucault, Michel. *The history of sexuality: An introduction*, volume I. Trans. Robert Hurley. New York: Vintage, 1990.

Foucault, Michel. The subject and power. Critical inquiry 8, no. 4 (1982): 777–795.

Heidegger, Martin. *The question concerning technology*. New York: Harper & Row, 1977.

Ihde, Don. *Technics and praxis: A philosophy of technology*. Vol. 24. Springer Science & Business Media, 2012.

Simons, Jon. *Foucault and the Political*. Psychology Press, 1995.