AI & SOCIETY (2020) 35:755–758
https://doi.org/10.1007/s00146-020-00947-7

CURMUDGEON CORNER

# The revelation of superintelligence

**Konrad Szocik[1] · Bartłomiej Tkacz[2] · Patryk Gulczyński[3]**

Received: 4 January 2020 / Accepted: 21 January 2020 / Published online: 8 February 2020


**Abstract**
The idea of superintelligence is a source of mainly philosophical and ethical considerations. Those considerations are rooted in the idea that an entity which is more intelligent than humans, may evolve in some point in the future. For obvious reasons, the superintelligence is considered as a kind of existential threat for humanity. In this essay, we discuss two ideas. One of them is the putative nature of future superintelligence which does not necessary need to be harmful for humanity. Our key idea states that the superintelligence does not need to assess its own survival as the highest value. As a kind of intelligence that is not biological, it is not clear what kind of attitude the superintelligent entity may evolve towards living organisms. Our second idea refers to the possible revelation of superintelligence. We assume that the self-revelation of such entity cannot be random. The metaphor of God as a superintelligence is introduced here as a helpful conceptual tool.

**Keywords** Superintelligence · Existential threat · Future studies · Thought experiment · revelation · Embodied mind · Ethics

## 1 Introduction

The idea of superintelligence should be, first of all, considered as a kind of the philosophical thought experiment which may be used as a good platform to study our moral intuitions or to create the hypothetical scenarios of the future human development. However, the probability that something like the superintelligence may ever happen, is appreciated by philosophers and futurists but often questioned by IT scholars. The idea that the intelligence as high as human intelligence or even higher may be realized beyond the human brain and body is philosophically appealing, ethically—and also theologically—challenging but technically almost implausible (Floridi 2016). Here, we want to discuss the challenge of the possible revelation of superintelligence in relation to its nature. When we understand the nature of superintelligence, we should find that the moment of its revelation is not a trivial issue. The challenge lies in the fact that—if the superintelligence is something more intelligent than we are—it should not be interpreted and understood in the anthropomorphic way. In some sense, humans are not prepared intellectually to understand and to predict decisions and behaviours of an entity which is more intelligent than them.

✉ Konrad Szocik
kszocik@wsiz.rzeszow.pl

Bartłomiej Tkacz
tkacz@kirp.pl

Patryk Gulczyński
w60059@student.wsiz.rzeszow.pl

[1] Department of Social Sciences, University of Information Technology and Management in Rzeszow, Sucharskiego 2 Street, 35-225 Rzeszów, Poland

[2] Member of the Board of the Polish National Bar of Attorneys in Warsaw, Warsaw, Poland

[3] University of Information Technology and Management in Rzeszow, Rzeszow, Poland

## 2 The nature of superintelligence

There are good reasons to assume that the superintelligence is a kind of entity which will exceed the boundaries typical for the human species. This is a kind of an intellectual challenge for humans, to think about the virtual mental capabilities of superintelligence in terms different than anthropomorphic.

First, let us assume that the superintelligence is a kind of intelligence which is able to understand, to know and to predict everything. There are no boundaries and no limits for its thinking, creativity and innovation. This idea

possesses serious consequences. In fact, it means that the superintelligence may solve all problems and find solutions for all theoretical and practical challenges. In contrast to the superintelligence, human intelligence is constrained both theoretically and practically. Theoretical constraints mean that human intelligence is not able to create intellectual solutions for particular issues. Such issues include, among others, the possibility of eternal life or human intergalactic journeys (Tegmark 2017). Humanity does not—as of yet—have any idea how to reach an eternity or how to send humans to another galaxy. Those issues are the kinds of issues which may remain always unsolvable for human intelligence. Another kind of constraints, practical constraints, involves issues which fit more or less to the human intellectual capacities. This kind of issues includes, for instance, a human mission to Mars. While this mission may be realized by humans in the next decades and, as a such, it does not go beyond the human technological capacities like the intergalactic travels, there are still serious practical constraints. The human intelligence is not able to offer solutions which will make that first human mission to Mars more cost effective, faster and safer for astronauts. While mission planners theoretically know the most effective and the most optimal solutions and strategies including artificial gravity in the spacecraft, terraformed Mars or hibernated astronauts, those theoretical considerations meet practical constraints. Human intelligence, in fact, is not able today to solve those practical constraints, and those optimal solutions do not go beyond the limit of speculation and conceptual models. Something similar is the case of global warming. While humans are able to predict the future climate changes and find the collateral causes, they either cannot invent and apply effective countermeasures, or quite simply do not see it as profitable as the fossil fuels industry. If said countermeasures were applied, however, they could offer a balance between a growing global human population, limited resources, environmental pollution and a sustainable environment. We can take for granted that superintelligence is able to find theoretical explanations and applicable solutions. This is the essence of superintelligence: theoretical considerations are identical with practical solutions because the superintelligence not only creates theoretical models—this is what humans can do—but it also finds the ways of their practical applications—which is what humans cannot do in many cases.

That correlation between the theoretical and practical contexts of intelligence has some ethical implications. Those ethical consequences are the starting point for the idea that the superintelligence is one of the most dangerous existential threats for humanity (Bostrom 2014; Brundage et al. 2018). While the idea of superintelligence does not imply that it must be or should be harmful for humans, no one is able to predict what kind of attitude towards humanity may be developed by the superintelligent being. However, there are strong reasons to assume that the attitude of superintelligence towards humanity will not be permanently good or bad but context-dependant. It may be a kind of a value-neutral ethical attitude which goes beyond the moral good and evil. The ethics of AI is an interesting issue. The first challenge arises what kind of ethical norms—if any—the superintelligence may have and apply. It should be asked, however, if superintelligence would even consider fixing human problems. The main problem is that we desire for our perceived problems to be fixed, while a superintelligent being might not even think that an idea of a problem actually exist, and that all human problems are the cause of their imperfect humanity. The second challenge is connected with the fact that superintelligence is a kind of a non-biological intelligence. There is no intelligent being at least as intelligent as humans which is not biological. Consequently, the high general intelligence is always connected with the body (Weigmann 2012; Maclure 2019). There is no doubt that morality and ethics have the biological roots and, as a such, they are integrally connected with body, sensuality and a capacity to feel pain and pleasure, just to mention a few main factors in the evolution of morality. Because the superintelligence means an entity that is not biological, no one is able to predict the possible evolution of moral "intuitions" and ethical norms. This issue is associated with the question of consciousness. We could even question whether or not superintelligence could exist without a moral compass, as even being uncaring for all life is hyper-nihilistic moral stance. It could be that, without a moral compass and evaluation between right and wrong, there could not exist the idea of will. While consciousness and intelligence are two distinct features, they always co-exist in the natural world—at least in regard to the living forms which possess a general intelligence (Maclure 2019). Consequently, we are not able to create a feasible model of decision-making of superintelligence in regard to our moral intuitions and ethical norms.

Paradoxically, this fact is for humans both good and bad. This is good because it means that we cannot predict any ethical determination appropriate for the superintelligence. It means that the superintelligence is not necessary by definition the most challenging existential threat for humanity. However, on the other side, the specificity of the superintelligence and an open-ended nature of its speculation opens space for all possible scenarios. None of which can be easily conjured by a human mind.

The issue is getting more and more complicated when one is trying to discuss the way in which the superintelligence may assess the value of the biological life. There are at work at least two possible scenarios. One of them states that the superintelligence cares for life in the same way as humans do. This is logically unreasonable anthropomorphic way

of thinking. The two following issues appear here. One of the issues is the mentioned fact that the superintelligence is not biological entity and, as a such, it does not share the common—in terms of homological or analogical similarities (Bergstrom and Dugatkin 2016)—biological evolutionary history. However, we could speculate that since the idea of superintelligence is non-human in nature, the possibility of the superintelligence being created in its infantile state by the humans and throughout time growing up to its ideal superintelligence could be described as something akin to be cybernetic evolution. Consequently, no one should expect that the superintelligence will assess the value of a biological life because the superintelligence does not share the biological life. The second issue is the fact that an attempt to assign a human-like morality to the superintelligence is a kind of a rationalization fallacy. That fallacy appears when one assumes that the human reason is the main and sufficient factor in the moral evolution (Buchanan and Powell 2018). This is not necessarily true because the evolution of morality is context-dependent and deeply rooted in the human biology and evolution which involves many non-rational factors such as biases, instincts and emotions, which all could be summed up in the idea of religion. The belief in supernatural had been around for centuries and always had been used as a sort of a multi-functional tool, capable of achieving anything through the ideas of reward and punishment. Religion was often used as a way to explain earthly phenomena; practice which was later adapted by the meta-physics, and followed by philosophy. Superintelligence would most likely have no need for religion as it would be god-like from a human point of view, thus it might have no need for moral compass. The main challenge is as follows: what kind of ethics may evolve in purely intellectual entity? There is no analogy in the natural world. While we have rationalistic ethical systems in the history of philosophy, all of them are developed by human philosophers. The rationalistic ethical system created by humans is not identical with the rationalistic ethical system created by a non-human intelligent system.

An alternative scenario states that the superintelligence does not care for the biological life. In this scenario, there is nothing valuable in the biological life itself. Here we should not apply our current ethical perspective which assigns to the biological life a high value—however, this is a species-dependant assessment. But, at least, the value of human life is almost an absolute value, and the humanitarian approach to the non-human animals is strongly developed at least in the Western world. There are no strong reasons to assume that the superintelligence should be obligated to value any living form—both human and nonhuman as well. The following question arises here. While the superintelligence extends the human intellectual and cognitive capacities, this is not clear if the pure intelligence—even if as advanced as ever possible—may create an idea of life as an inherent value. No human is able to solve that dilemma due to the fact that we do not know any ethical system which is created by entities that are not human and who do not share the common evolutionary history.

## 3 The revelation of superintelligence

Here, we want to discuss one particular issue within the philosophical study on Artificial Intelligence: its "revelation". We assume that one of the consequences of the specificity of the superintelligence is the fact that the day when it may appear, it may be a kind of mystery. There are good reasons to assume that the superintelligence will be not necessary prone to self-reveal its existence to humans. That assumption is a logical consequence of the super intelligent nature of that entity. However, there is a vicious circle of human thinking that is hard to avoid: we assume here a kind of a survival instinct which is supposed to be a capacity of the superintelligence. But such kind of instinct is not a feature of the intelligence itself but this is a feature of the living organisms. From the fact that the superintelligence is, by definition, the most intelligent entity, it does not mean that such entity would have any kind of survival instinct. This is rather our anthropomorphic cognitive bias which lies in the fact that we are biological living forms and, as a such, we possess such kind of instinct which we share with all other living forms. This is a hard subject to think about, an autonomous and intelligent entity which may not possess such instinct. This assumption is the basic one for the idea that the superintelligence may cause an existential threat for humanity. Superintelligence may decide to kill humans when it finds that humanity is some kind of danger for its existence. The survival instinct is connected with the self-assessment of the value of the superintelligence's own existence. This is also not clear whether the superintelligence may perceive its own existence as valuable and, as a such, worthy of protection. While we can assume that the definition of the superintelligence includes the idea that the superintelligence should know how to survive and protect its existence, from that definition does not follow the fact that the superintelligence wants to and should do that. It seems that the fact of being a biological living form equipped in emotions and feelings is a crucial factor to have such capability. Because the superintelligence is not a biological living form, this is not clear if we should expect to find the superintelligent entity which will take care of its own survival. It is worth keeping in mind the mentioned fact that knowing how does not imply the duty to do that. Another possible scenario when the superintelligence may decide to kill humanity is the possibility that the superintelligence finds a human life useless—but not hazardous to the superintelligence.

The way of self-manifestation of the superintelligence may be at least partially affected by the mentioned survival instinct. If we assume that the superintelligence possesses such kind of instinct like the biological living forms do—independently on the fact that the superintelligence is not a living form—we may expect that the superintelligence will care for the way, time and place of its self-revelation. Here once again the theological metaphor may be at work. The revelation of the superintelligence may be like the biblical God's revelation. According to the biblical story, God is the intelligence which goes beyond the human intelligence, and he reveals himself in a non-random way to non-random people.

Let's apply that biblical story about God's revelation to the possible revelation of the superintelligence. The question arises not when the superintelligence may appear and when it may decide to let us know about it. The question is that the superintelligence—like a biblical God—may decide to self-reveal its existence to particular, non-random people. It is hard to speculate about the virtual intents and motivations—if any—of the superintelligence. However, we can take for granted that the superintelligence equipped in a kind of the survival instinct should care for time and place of its self-revelation. One of the crucial factors is an access to resources and sources of energy. The superintelligence may wait for the proper time when it will be sure that it is safe in the sense of its energetic safety.

These are, however, speculations, and as such do not reflect reality. While we may have an idea of what superintelligence is, where it will come from and when it will reveal itself, they all end up being just ideas, which raises even more important question. If reveal of superintelligence in is inevitable, does it really matter what we think of it since as we came to the conclusion, our idea of such concept does accurately reflect reality. The metaphor of God as a superintelligence may be a useful starting point for an attempt to define the essence and nature of super -intelligent entity in anthropomorphic terms.

**Curmudgeon Corner** is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? –Editor.



## References

Bergstrom CT, Dugatkin LA (2016) Evolution, 2nd edn. W. W. Norton & Company, New York

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Brundage M et al. (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, Open AI

Buchanan, A., R. Powell. *The Evolution of Moral Progress. A Biocultural Theory*, Oxford University Press, 2018

Floridi L (2016) Should we be afraid of AI? Aeon, 9 May, 2016. https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Maclure J (2019) The new AI spring: a deflationary view. AI & Soc. https://doi.org/10.1007/s00146-019-00912-z

Tegmark M (2017) Life 3.0: being human in the age of artificial intelligence. Alfred A. Knopf, New York

Weigmann K (2012) Does intelligence require a body? The growing discipline of embodied cognition suggests that to understand the world, we must experience the world. EMBO Rep 13(12):1066–1069