Müller, Vincent C. (2016), 'Editorial: Risks of artificial intelligence', in Vincent C. Müller (ed.), *Risks of general intelligence* (London: CRC Press - Chapman & Hall), 1-8.
http://www.sophia.de
http://orcid.org/0000-0002-4144-4957

# Editorial: Risks of Artificial Intelligence

*Vincent C. Müller*

Anatolia College/ACT
www.sophia.de

*Abstract:* If the intelligence of artificial systems were to surpass that of humans significantly, this would constitute a significant risk for humanity. Time has come to consider these issues, and this consideration must include progress in AI as much as insights from the theory of AI. The papers in this volume try to make cautious headway in setting the problem, evaluating predictions on the future of AI, proposing ways to ensure that AI systems will be beneficial to humans – and critically evaluating such proposals.

## 1. Introduction: Risk of AI

This is the first edited volume on the risks of AI and it originates from the first conference on risks of AI (AGI-Impacts). Following that conference we published a volume of selected papers in the *Journal of Experimental and Theoretical Artificial Intelligence* (JETAI) – (see Müller, 2014). Our volume generated significant interest: It had ca. 20.000 paper downloads from the JETAI site alone in the first year; three of the top-five downloaded papers in JETAI are now from our volume. As a result, the publishers suggested turning the journal volume into a book and adding some recent material, so this is what you are holding in your hands.

The notion that AI might generate existential risk to humanity has gained currency since we published the journal volume: Nick Bostrom's book *Superintelligence: Paths, Dangers, Strategies* has come out (Bostrom, 2014), well known public intellectuals like Stephen Hawking and Stuart Russell have published warning notes in the general press (Hawking, Russell, Tegmark, & Wilczek, 2014) and a host of media publications have followed. The idea of existential risk in its Hollywood format is 'the machines will take over and kill us all' – this fear obviously strikes a cord. The spread of this fear has generated significant concern amongst academics in AI. One indication is that the current and past presidents of the AAAI (that most significant academic AI association) wrote a short statement to the effect that "AI doomsday scenarios belong more in the realm of science fiction than science fact", while they also "urge our colleagues in industry and academia to join us in identifying and studying these risks" (Ditterich & Horowitz, 2015). Recently, some efforts were made to outlined the research agenda

for AI that is beneficial for humanity, e.g. in the new 'Future of Life Institute' (Russell, Dewey, & Tegmark, 2015) and the renamed 'Machine Intelligence Research Institute', MIRI (Soares & Fallenstein, 2014).

While the traditional concerns in the philosophy and theory of AI have focused on the prospects of AI, its relation to cognitive science and its fundamental problems (see the www.pt-ai.org conference series), we now see an increasing focus on matters of risk and ethics. But what is the general idea of this shift?

## 2. The risks of artificial intelligence

The notion of an agent with general intelligent ability is the original driving vision of AI research (see McCarthy, Minsky, Rochester, & Shannon, 1955) and dominates much of its public image – while nearly all actual current work in AI is on specialised technology, far removed from such a general ability, and often without use of the term "artificial intelligence".

The move from AI to risk is relatively easy: There is no reason to think that the level of human intelligence is anything special in the space of possibilities – it is easy to imagine natural or artificial intelligent agents that are vastly superior to us. There also seem to be reasons to think that the development of artificial intelligence is accelerating, together with related technologies, and that the invention of intelligent machines itself would further accelerate this development, thus constituting an 'argument from acceleration' for the hypothesis that some disruptive transformation will occur. So, if one thinks of intelligence as a quantifiable unit then this acceleration will continue and move past the (small) space that marks the intelligence of humans. So, we will reach 'superintelligence', which Bostrom tentatively defines as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014, p. 22). In an classic passage, Good has speculated that "the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" (Good, 1965, section 2). So, there is the risk that 'the machines take over' and this loss of control is a significant risk, perhaps an existential risk for humanity (for a survey, see Sotala & Yampolskiy, 2013).

The discussion of risk is *not* dependent on the view that AI is now on a successful path towards superintelligence – though it gains urgency if such 'success' is a non-negligible possibility in the coming decades. It also gains urgency if the stakes are set high, even up to human extinction. If the stakes are so high as to include extinction of humankind, even a fairly small possibility of a disastrous outcome (say, 3%) is entirely sufficient to motivate the research. Consider that if there were a 3% possibility that a plane you are about to board will crash: That would be sufficient motivation for getting off. The utility at stake in scientific or philosophical research is usually quite a bit

lower. It appears that the outcome of superintelligence is more likely to be extreme: either extremely bad or extremely good for humanity.

As it happens, according to our recent research, the estimation of technical experts is that by 2050 the probability of high-level machine intelligence (that surpasses human ability in nearly all respects) goes beyond the 50% mark, i.e. it becomes more probable than not (Müller & Bostrom, forthcoming 2014). 2050 is also the mark that 'RoboCup' set itself for fielding a robot team that can beat the human football world champions (actually an aim that does not make much sense).

## 3. The papers

The paper by Omohundro (2) introduces the problem of risk and the author presses his point that even an innocuous artificial agent, like one programmed to win chess games, can very easily turn into a serious threat for humans, e.g. if it starts acquiring resources to accomplish its goals: "The seemingly harmless chess goal therefore motivates harmful activities like breaking into computers and robbing banks." (section 4.2). He suggests that we need formal methods that provide proofs of safe systems, a "Safe-AI Scaffolding Strategy".

The two following papers deal with prediction of coming success in AI: Armstrong/Sotala/O'Heigeartaigh (3) propose a decomposition schema to compare predictions on the future of AI and then test five famous predictions, from the Dartmouth Conference, Dreyfus, Searle, Kurzweil and Omohundro – with the result that they are poor, especially the optimistic ones. T. Goertzel (4) argues that while most progress in AI so far has been 'narrow' technical AI, the next stage of development of AI, for at least the next decade and more likely for the next twenty-five years, will be increasingly dependent on contributions from strong-AI.

From here, we go into the proposals on how to achieve safer and ethical general AI. In paper (5) Brundage investigates the general limitations of the approach to supply an AI with a 'machine ethics', and finds them both serious and deeply rooted in the nature of ethics itself. Yampolskiy (6) investigates which utility functions we might want to implement in artificial agents and particularly how we might prevent them from finding simple but counterproductive self-satisfaction solutions. B. Goertzel explains (7) how his "Goal-Oriented LEarning Meta-Architecture" (GOLEM) may be capable of preserving its initial – benevolent – goals while learning and improving its general intelligence. Potapov and Rodinov (8) outline an approach to machine ethics in AIXI that is not based on 'rewards' (utility) but on learning 'values' from more 'mature' systems. Kornai argues (9) that Alan Gewirth's dialectical argument, a version of classic Kantian ethical rationalism, shows how an artificial agent with a certain level of rationality and autonomy will necessarily come to understand what is moral. Kornai thus denies what Bostrom calls the 'orthogonality thesis' (Bostrom, 2012), namely that ethical motivation and intelligence are independent or 'orthogonal'.

Last but not least, Sandberg (10), looks at the special case of general AI via whole brain emulation, in particular, he considers the ethical status of such an emulation: Would the emulation (e.g. of a lab animal's brain) have the ability to suffer, would it have rights?

In our new contributions for this volume Dewey (11) investigates strategies to mitigate the risk from a fast takeoff to superintelligence in more detail. Bishop (12) takes a different line and argues that there is no good reason to worry about existential risk from AI, but that we should rather be concerned about risks that we know are coming – such as the military use of AI. Like many people working in AI, Bishop remains unimpressed by the discussion about risks of superintelligence because he thinks there are principled reasons why machines will not reach these abilities: they will lack phenomenal consciousness, understanding and insight.

## 4. Outlook: Ethics of AI and Existential Risks of AI

These last two contributions are perhaps characteristic of a divide opening up in the debates between the 'normal ethics' side that stresses the challenges for AI and a 'existential risks' side that stresses the big challenges for humanity. In the 'existential risks' tradition, the traditionally central issues of consciousness, intentionality and mental content are literally dispensed with in a footnote (Bostrom, 2014, fn. 2 to p. 22) and embodied cognition is not mentioned at all, like any other cognitive science.

I tend to think that both extremes are unlikely to be fruitful: to stick to the traditional problems and to ignore them. It is unlikely that nothing can be learned about the long-term future of AI from the critics of AI, and it is equally unlikely that nothing can be learned about that future from the actual success of AI. So, while Dreyfus is right to say that the history of AI is full of 'first step fallacies' that are similar to claiming that "the first monkey that climbed a tree was making progress towards landing on the moon" (Hubert L. Dreyfus, 2012, p. 92); Bostrom is right to say "From the fact that some individuals have overpredicted artificial intelligence in the past, however, it does not follow that AI is impossible or will never be developed." (Bostrom, 2014, p. 4).

As I noted in my earlier editorial (Müller, 2014) (which shares some text with this one), the term 'singularity' is now pretty much discredited in academic circles – with the notable exception of (Chalmers, 2010) and the ensuing debate. It is characteristic that the only paper here that uses it (Bishop) is critical of the notion. Singularity is associated with ideological techno-optimism, trans-humanism and predictions like those of Ray Kurzweil (esp. Kurzweil, 2005; more recently Kurzweil, 2012) that ignore the deep difficulties and risks of AI, e.g. by equating intelligence and computing power. What was the 'Singularity Institute' is now called the 'Machine Intelligence Research Institute' (MIRI). 'Singularity' is on its way towards becoming, literally, the trademark of a particular ideology, without academic credentials.

The most important thing between 'existential risks' and 'normal ethics' is to realise that both sides could be wrong: It might be that superintelligence will never be developed (Bishop) and it might be that it will likely be developed (Bostrom) – but if both are possible then we would do well to look into the consequences (Bostrom), while taking the arguments about constraints (Bishop) into account. We need to talk. When we do that we will realise there is much to learn from the 'other' side:

The problem of identifying the risks of general AI and even controlling them before one knows what form or forms that general AI might take, is rather formidable. To make things worse, we don't know when the move from fairly good AI to a human and then superintelligent level might occur (if at all) and whether it will be slow enough to prepare or perhaps quite rapid – it is often referred to as an 'explosion' (see Dewey in this volume). As we have seen above, one might try to mitigate the risks from a superintelligent goal-directed agent by making it 'friendly' (see e.g. Muehlhauser & Bostrom, 2014), by 'controlling' or 'boxing' it or just by trusting that any superintelligent agent would be already 'good'. All these approaches make rather substantial assumptions about the nature of the problem, however; for instance, they assume that superintelligence takes the form of an *agent* with goals, rather like us. It is even doubtful that some assumptions about agency are consistent: Can an agent have goals (rather than just a technical 'utility function') without having the ability for pain and pleasure, i.e. phenomenal experience? Of course, it is conceivable that superintelligence will take very different forms, e.g. with no individuality or no goals at all, perhaps because it lacks conscious experience, desires, intentional states or an embodiment. Notoriously, classical critics of AI (Hubert L Dreyfus, 1992; Searle, 1980) and more recent cognitive science have provided arguments that indicate which directions AI is unlikely to take, and full agency is among them (Clark, 2008; Haugeland, 1995; Pfeifer & Bongard, 2007; Varela, Thompson, & Rosch, 1991).

Of course, superintelligence may constitute a risk without being an agent, but what do we really know about it, then? Even if intelligence is not deeply mysterious and fundamentally incomparable, as some people claim, it is surely not a simple property with a one-dimensional metric either. So, just saying that a general artificial intelligence is, well, 'intelligent', does not tell us much: As Yudkowsky urges, "One should resist the temptation to spread quantifiers over all possible minds." (2012, p. 186) – if that is true, the temptation to say anything about the even larger set of 'possible intelligent systems' is also to be resisted. Certainly, we should say what we mean by 'intelligent' when we claim that 'superintelligence' is coming and would constitute existential risk.

There is a serious question whether rigorous work is even possible at this point, given that we are speculating about the risks from something about which we know very little. The current state of AI is not sufficiently specific to limit that space of possibilities enough. To make matters worse, the object of our study may be *more* intelli-

gent than us, perhaps far more intelligent, which seems to imply (though this needs clarification) that even if we were to know a lot about it, its ways must ultimately remain unfathomable and uncontrollable to us mere humans.

Given these formidable obstacles our efforts are at danger to look more like theological speculation or ideological fervour than like science or analytic philosophy. We are walking a fine line and have to tread very carefully. The papers in this volume are trying to make some headway in this difficult territory since we remain convinced that cautious progress is better than rushing headlong into the dark.

## References

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines, 22*(2 - special issue 'Philosophy of AI' ed. Vincent C. Müller), 71-85.

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

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9-10), 7-65.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Ditterich, T., & Horowitz, E. (2015). Benefits and risks of artificial intelligence. *medium.com.* Retrieved 23.01.2015, from https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason* (2 ed.). Cambridge, Mass.: MIT Press.

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines, 22*(2 - special issue "Philosophy of AI" ed. Vincent C. Müller), 87-99.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Ruminoff (Eds.), *Advances in Computers* (Vol. 6, pp. 31-88). New York & London: Academic Press.

Haugeland, J. (1995). Mind embodied and embedded. *Acta Philosophica Fennica, 58*, 233–267.

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014). Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough? *The Independent, 01.05.2014*.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. London: Viking.

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. New York: Viking.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Retrieved October 2006, from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Muehlhauser, L., & Bostrom, N. (2014). Why we need friendly AI. *Think, 13*(36), 41-47. doi: 10.1017/S1477175613000316

Müller, V. C. (2014). Editorial: Risks of general artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence, 26*(3 - Special issue 'Risks of General Artificial Intelligence', ed. V. Müller), 1-5.

Müller, V. C., & Bostrom, N. (forthcoming 2014). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence*. Berlin: Springer.

Pfeifer, R., & Bongard, J. (2007). *How the body shapes the way we think: A new view of intelligence*. Cambridge, Mass.: MIT Press.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. 2015, from http://futureoflife.org/static/data/documents/research_priorities.pdf

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417-457.

Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report, 2014*(8).

Sotala, K., & Yampolskiy, R. V. (2013). Responses to Catastrophic AGI Risk: A Survey. *MIRI Technical Reports, 2013*(2).

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. Cambridge, Mass.: MIT Press.

Yudkowsky, E. (2012). Friendly artificial intelligence. In A. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 181-194). Berlin: Springer.