Journal of Experimental & Theoretical Artificial Intelligence, 2014
Vol. 26, No. 3, 297–301, http://dx.doi.org/10.1080/0952813X.2014.895110

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Risks of general artificial intelligence

1. The conference

The papers in this special volume of the *Journal of Experimental and Theoretical Artificial Intelligence* are the outcome of a conference on the 'Impacts and Risks of Artificial General Intelligence' (AGI-Impacts) that took place at the University of Oxford, St Anne's College, on 10 and 11 December 2012 – jointly with the fifth annual conference on 'Artificial General Intelligence' (AGI-12). The conference was organised by the Future of Humanity Institute at Oxford: academically by Nick Bostrom and myself, with support from research fellows Stuart Armstrong, Toby Ord, Anders Sandberg and more members of the program committee; organisationally by Seán Ó hÉigeartaigh, with support from Alexandre Erler, Daniel Dewey, Stuart Armstrong and others.

We are grateful to the Artificial General Intelligence (AGI) community for the openness to these issues of security, as shown by the fact that they initiated this connection and that the vast majority of the ca. 150 participants of the main AGI meeting also attended our AGI-Impacts event. Last but not least, we want to thank the 'European Network for Cognitive Systems, Interaction and Robotics' and the organisation 'Saving Homo Sapiens' for sponsoring the event.

2. The risks of general artificial intelligence

The notion of an agent with general intelligent ability is surely the original driving vision of artificial intelligence (AI) research (see McCarthy, Minsky, Rochester, & Shannon, 1955) and dominates much of its public image, but 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 AI. Some researchers who wish to return to the original vision have formed the AGI community that met at the sister conference in Oxford (http://agi-conference.org/2012). People in the AGI community say that the original vision of human-level general intelligence was a fine one and that time is ripe for this return to the roots because the extant approaches are in principle sufficient to achieve it in the foreseeable future. (For a general research concept, see Adams et al., 2012.)

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 AI 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 (see Eden, Moor, Søraker, & Steinhart, 2012, p. 2). One possibility is 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

This article was originally published with errors. This version has been corrected. Please see corrigendum (http://dx.doi.org10.1080/0952813x.2014.963389).

keep it under control' (Good, 1965, Section 2). If high-level AI occurs, this will have a significant impact on humanity, especially on the ability of humans to control their fate on Earth. 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 AGI is on a successful path towards human-level AI – 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, even a fairly small possibility (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 to be gained from scientific or philosophical research is usually quite a bit lower.

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, in press) – on whether one can trust such estimations, see Armstrong, Sotala & Ó hÉigeartaigh, this volume.

3. The papers

We called for papers with this description:

> The conference explores questions such as: How can we best predict the impact of future intelligent and superintelligent machines? How can we combine ideas from computer science, mathematics and philosophy to best estimate this impact? What will be the impacts of AGI on the world? Which directions of research should be most explored, and which should be de-emphasized or avoided? What can we do to best ensure scientific rigour in this non-experimental academic field? What are the best ideas and methods for ensuring both safety and predictability of advanced AI systems? Can we lay the foundations to a field of rigorous study of realistic AGI control methods that lead to implementable security protocols?

We had 39 submissions by the deadline, of which we accepted 11 (28%) after double-blind review. The event also featured a keynote talk by Bruce Schneier and one by Steve Omohundro. Of the 11 papers and 2 keynotes, 9 made a timely submission for this volume and survived a further review by the editor, including a pass through automated plagiarism software (very useful to catch sloppy referencing and self-plagiarism/recycling). It turned out that the 'risks' took precedence over the general 'impacts' in the submissions, and we thus dropped the term 'impacts' from the title of this volume.

Schneier presented the theory from his new book *Liars and Outliers: Enabling the Trust that Society Needs to Thrive* (Schneier, 2012) that security needs to mature to a wider discipline, which crucially relies on establishing and maintaining trust, a trust that is undermined by many, including state agents. Margaret Boden, Nick Bostrom and Angelo Cangelosi spoke at the main AGI conference. We have video interviews with Boden, Schneier and Aaron Sloman on the conference site at http://www.winterintelligence.org/oxford2012/agi-impacts/videos/.

The paper by Omohundro (2014) 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: Armstrong, Sotala & Ó hÉigeartaigh (2014) 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 (2014b) 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 25 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 the fifth paper Brundage (2014) 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 (2014) 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 (2014a) explains how his 'Goal-Oriented Learning Meta-Architecture' may be capable of preserving its initial – benevolent – goals while learning and improving its general intelligence. Potapov and Rodinov (2014) outline an approach to machine ethics in AIXI that is not based on 'rewards' (utility) but on learning 'values' from more 'mature' systems. AIXI is a Bayesian optimality concept for reinforcement learning agents in unknown environments (see Hutter, 2005). Kornai (2014) argues 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.

Last but not least, Sandberg (2014) 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?

4. The outlook

Perhaps it may be permitted to add two notes, from the perspective of the editor:

A note on *terminology*: It is characteristic that none of the authors in this volume uses the term 'singularity' to characterise future development of AI – in fact, we had only a single paper submission using this word in the title or subtitle. People prefer other, more specific terms like 'intelligence explosion', 'AGI', 'superintelligence', 'acceleration', etc. It would appear 'singularity' is now pretty much discredited in academic circles – with the notable exception of Chalmers (2010) and the ensuing debate. The discussions about singularity are generally characterised by conviction and fervour, which support amateurism and vitriolic exchanges – even in academically respectable publications like the comments in Eden et al. (2012). 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.

A note on *methodology*: Of course, 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'. 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. 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 (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). 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? If not, then an agent with goals is also a moral patient and we have to treat it ethically.

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. So, just saying that a general AI 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. 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* intelligent 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 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 nothing – and more work in this direction will be available in Bostrom (in press) – but caution must remain our primary guide.

References

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., & Sowa, J. F. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine*, *33*, 25–42.

Armstrong, S., Sotala, K., & Ó hÉigeartaigh, S. S. (2014). The errors, insights and lessons of famous Al predictions – and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 317–342.

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

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 355–372.

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

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

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

Eden, A., Moor, J. H., Søraker, J. H., & Steinhart, E. (Eds.). (2012). *Singularity hypotheses: A scientific and philosophical assessment*. Berlin: Springer.

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, NY: Academic Press.

Goertzel, B. (2014a). GOLEM: towards an AGI meta-architecture enabling both goal preservation and radical self-improvement. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 391–403.

Goertzel, T. (2014b). The path to more general artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 343–354.

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

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer.

Kornai, A. (2014). Bounding the impact of AGI. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 417–438.

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, NY: 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*, 41–47. doi:10.1017/S1477175613000316.

Müller, V. C., & Bostrom, N. (in press). Future progress in artificial intelligence: A poll among experts. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence*. Berlin: Springer.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 303–315.

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

Potapov, A. & Rodionov, S. (2014). Universal empathy and ethical bias for artificial general intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 405–416.

Sandberg, A. (2014). Ethics of brain emulations. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 439–457.

Schneier, B. (2012). *Liars and outliers: Enabling the trust that society needs to thrive*. New York, NY: Wiley.

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

Sotala, K., & Yampolskiy, R. V. (2013). *Responses to catastrophic AGI risk: A survey* (Technical Reports, 2013(2)). Berkeley, CA: Machine Intelligence Research Institute.

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

Yampolskiy, R. V. (2014). Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*, 373–389.

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.

Vincent C. Müller
Future of Humanity Institute & Oxford Martin School,
University of Oxford, Oxford, UK
Anatolia College/ACT, Thessaloniki, Greece
www.sophia.de; Email: vincent.mueller@philosophy.ox.ac.uk