Guest Editors' Introduction to the Special Issue on "Superintelligence"

The concept of superintelligence was developed only a little after the birth of the field of artificial intelligence, and it has been a source of persistent intrigue ever since. Alan Turing himself toyed with the idea of human-level intelligence: "If a machine can think, it might think more intelligently than we do, and then where should we be?"[1] In 1965, I.J. Good, a former colleague of Turing's, considered what would happen if a machine could effectively redesign itself.[2] This, he argued, could lead to what we would now call a superintelligence: a system that "greatly exceeds the cognitive performance of humans in virtually all domains of interest".[3]

In some ways, our understanding of how to deal with these problems has advanced little since then. Although the field of artificial intelligence has advanced substantially, it is quite unclear by what pathway superintelligence may be reached. Given that technological forecasting is covered in such a haze, we cannot say that superintelligent AI will come soon, but neither can we be assured that it will be far away. It would be similarly complacent to claim to know with confidence whether such a system will be beneficial or harmful by default. It is troubling that we still find ourselves so uncertain about these 'crucial considerations'[4]: the emergence of 'human' intelligence proved a watershed in the history of the earth (certainly in our history), and the prospective development of superintelligence is unlikely to be any smaller in its impact and ramifications. Now may be (and is on expectation), a critical time to think more, so that we can see a sharper outline of this situation, and formulate plan for managing it.

Over the last decade, a range of academics have finally begun to respond to this challenge in an organized way. Many core philosophical issues have been charted, and technical AI safety research is now an emerging field;[5] there is also a young but ambitious research agenda exploring the geopolitical impacts and governance challenges surrounding the eventual deployment of superintelligent systems.[6] Some of this research can find a natural home in the fields of computer science, political science, or philosophy. Much, however, cannot. The considerations in evaluating and planning for superintelligence often cut across the practical and the philosophical, the technical and the non-technical, riding across several academic disciplines such that the most important work will often have no natural home in any of them. So the purpose of this Special Issue is to collect some of these essays, that use a full range of tools to evaluate and plan for superintelligence.

We hope that this will generate insights and debates that can help us get a better handle on this important topic--to enable us to undertake the conceptual, technological and societal innovations that will make superintelligence beneficial for the world.

The contributions to this issue can be coarsely separated into two baskets. Four of the contributions primarily focus on improving our understanding of the strategic landscape: they characterise the development of superintelligence, and its potential consequences. The remaining three chart a path through this landscape: they argue for specific kinds of research in order to make beneficial outcomes more likely.

In 'Superintelligence as a cause or cure for risks of astronomical suffering', **Kaj Sotala & Lukas Gloor** outline a new category of "suffering risks" ('s-risks'), in which astronomical suffering occurs on an astronomical scale. They propose that such risks may be of comparable severity and probability as extinction risks, and survey some of the ways that superintelligent AI might either bring about or relieve these kinds of risks, and some of the ways that further theoretical work could affect these.

In 'Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence', **Michael Batin, Alexey Turchin, Sergey Markov, Alisa Zhila, David Denkenberger** consider how steadily advancing AI could be used to extend the human lifespan. They offer an extensive survey of presently ongoing and potential future AI applications to anti-aging research, at three stages of development--narrow AI, AGI, and superintelligence, finding that medical-focused superintelligence might help humans to achieve 'longevity escape velocity'.

In 'Modeling and Interpreting Expert Disagreement About Artificial Superintelligence', **Seth Baum, Anthony Barrett, and Roman Yampolskiy** consider how we are to deal with persistent, pervasive expert disagreement about the risks posed by superintelligence. They describe a 'ASI-PATH' fault-tree model, and use it

[1] Turing, Alan. "Can digital computers think?(1951)." B. Jack Copeland (2004): 476.

[2] Good, I. J. "Speculations Concerning the First Ultraintelligent Machine*." Edited by Franz L. Alt and Moris Rubinoff. Advances in Computers 6 (1965): 31–88.

[3] Bostrom, 2014: 25.

[4] Bostrom, Nick. 2014. *Crucial considerations and wise philanthropy.*

[5] See agendas for this work at [Taylor, Jessica, et al. "Alignment for advanced machine learning systems." Machine Intelligence Research Institute (2016).] and [Amodei, Dario, et al. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).], and collections that include some of this work (as well as some other kinds) at https://futureoflife.org/ai-safety-research/ and http://effective-altruism.com/ea/1iu/2018_ai_safety_literature_review_and_charity/.

[6] Recent work in this area is compiled at http://www.allandafoe.com/aireadings

to chart points of disagreement between Nick Bostrom and Ben Goertzel, over the viability of catastrophic risks from superintelligence. They show how this model can assist with weighing the importance of different considerations, and can help with prioritization of superintelligence risk management strategies.

David Jilk, in 'Conceptual-Linguistic Superintelligence' reviews the ingredients and dynamics of an 'intelligence explosion', arguing that any AI system capable of sustaining such an intelligence explosion must have a 'conceptual-linguistic' faculty with functional similarity to that found in humans.

The papers that proposed future research were quite complementary to one another. Each of the three proposed a kind of research that could draw on different kinds of expertise to the others.

Gopal Sarma & Nick Hay, in 'Mammalian Value Systems', seek to bring fresh insights from other academic disciplines to bear on the problem of aligning AI goals with human values. They argue that what we call human values can be decomposed into (1) mammalian values, (2) human cognition, (3) human social and cultural evolution. They further argue that having more detailed prior information on the structures of human values may enable AI agents to infer these values from fewer examples, and advocate, on this basis, for greater research on mammalian values.

In their second submission, 'Robust computer Algebra, Theorem Proving and Oracle AI**', Sarma & Hay** provide another concrete avenue for AI safety research. They identify 'computer algebra systems' (CAS) as primitive examples of domain-specific oracles.; By charting efforts to integrate such computer algebra systems with theorem provers, they lay out a concrete set of encountered problems and considerations relevant to the 'provable safety' of eventual superintelligent 'Oracle AI'.

In 'The Technological Landscape Affecting Artificial General Intelligence and the Importance of Nanoscale Neural Probes', **Daniel Eth** argues that the development of nanoscale neural probes could substantially increase the likelihood that whole brain emulations are the first kind of AGI developed (as opposed to 'de novo' AI and neuromorphic AI). He argues as a result that it is desirable for research effort to be dedicated to accelerating their development.

Although the study of superintelligence has resurged in the last decade, it is still at a relatively early stage of maturity. It is one of the most exciting--and plausibly one of the most important--research areas of our time. As guest editors, we hope to have collected some work that has shone a small light on some of the problem of what to do about superintelligence. We are grateful to all authors for their contributions to this issue , and for their broader work exploring this critical topic. We also give special thanks to Prof. Matjaz Gams, Editor-in-chief of Informatica, for his support in composing this special issue.

Ryan Carey

Matthijs Maas

Nell Watson

Roman Yampolskiy