Institution: University of Oxford		
Unit of Assessment: 30 - Philosophy		
Title of case study: AI: Alignment, Policy and Governance		
Period when the underpinning research was undertaken: 1 January 2013–31 December 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Professor Nick Bostrom	Professor and Director, FHI	01 Oct 2012 – present
Professor Allan Dafoe	Associate Professor and Director, Centre for the Gov of AI	04 Sept 2018 – present
Dr Eric Drexler	Sr. Research Fellow	01 March 2018 – present
Dr Stuart Armstrong	Sr. Research Fellow	01 Aug 2015 – present
Dr Owen Cotton-Barratt	Sr. Research Fellow	01 Aug 2015 – present
Dr Owain Evans	Research Fellow	01 Feb 2017 - 22 Nov 2019
Miles Brundage	Research Fellow	25 July 2016 – 18 Aug 2018
Ben Garfinkel	Research Fellow	08 Oct 2018 – present
Period when the claimed impact occurred: 2014 to 31 December 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact (indicative maximum 100 words)

Professor Nick Bostrom's academic research on the dangers of highly advanced artificial intelligence was summarised in his New York Times bestseller *Superintelligence: paths, dangers, strategies* in 2014. The book has sold over 250,000 copies and has been disseminated via a variety of media outlets to over 7,000,000 people online. The impact of this publication, combined with work with researchers at the Future of Humanity Institute (FHI), led to changes in public understanding of risks posed by AI, and new policy about future treatment of AI. It has also led to new commercial and non-commercial policy on safety standards in AI development and deployment as well as the introduction of AI safety as a focus area for philanthropic funding.

2. Underpinning research (indicative maximum 500 words)

Led by and including Prof Bostrom, researchers at the University of Oxford's Future of Humanity Institute (FHI) investigated the nature of superintelligent AI systems, how they might help or harm society, and how we should respond.

Change in understanding of the 'control problem'

Bostrom's 2014 book *Superintelligence* [A] has improved the understanding of the 'control problem' (how can we create controls so that machine superintelligence will be beneficial instead of harmful to humanity?). Prior to the book, there was a distinct lack of rigorous mainstream and academic exploration of the implications of advances in AI and specifically the development of Artificial General Intelligence (AGI). *Superintelligence* served to consolidate arguments around the risk of such technologies. It developed many of the core concepts which have come to lay the foundation of the AI alignment field, like the orthogonality thesis (the assertion that capability and goals of an agent are independent of one another), and the examination of different types of possible superintelligent systems (e.g., Oracles, tools), as well as their associated risks.

In addition, whereas AI was previously seen primarily as a computer science discipline, *Superintelligence* raised considerations which demanded further engagement from disciplines

ranging from philosophy (e.g., the moral status of digital minds) and political science (e.g., how to treat international race dynamics in developing powerful AI; see also [B]) to economics (e.g., whether an intelligence explosion would lead to Malthusian conditions for the large majority of agents).

Subsequent research by Professor Bostrom and researchers at the FHI has built on this platform, focussing on techniques for building safer artificially intelligent systems. Work has included both theoretical (such as models of causal influence, and the limitations of value learning, e.g. [F]) and experimental (such as training deep learning models to decompose complex tasks, and to be more robust to large errors) aspects.

Centre for the Governance of AI
Established in 2018 and led by Professor Dafoe, this new research centre housed at FHI focuses on the political challenges arising from advanced AI by conducting research on important and neglected issues of AI governance, and advising decision makers on this research through policy engagement. For example, the 2020 report *The Windfall Clause: Distributing the Benefits of AI for the Common Good* [D] picks up and extends one of the economic concepts raised in *Superintelligence.*

Preventing the malicious use of artificial intelligence
A report involving several FHI researchers in partnership with NGOs and other academic partners [E] distilled findings from a workshop held in 2017, as well as subsequent research from the authors, to explore possible risks to security posed by malicious applications of AI in the digital, physical, and political domains, and mapped out a research agenda for further work in addressing such risks.

3. References to the research (indicative maximum of six references)

A. [Authored Book, available on request] Bostrom N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press. ISBN: 9780199678112
B. [Journal Article, listed in REF2] Bostrom, N. (2017). "Strategic implications of openness in AI development." *Global Policy*, 8(2), 135-148. DOI: 1111/1758-5899.12403
C. [Chapter, listed in REF2] Bostrom N, Dafoe A, Flynn C. 'Public Policy and Superintelligent AI: A Vector Field Approach'. in Liao, S.M. (ed.): *Ethics of Artificial Intelligence*. Oxford University Press, 2020. ISBN: 9780190905040
D. [Conference Contribution] O'Keefe, C., Cihon, P., Flynn, C., Garfinkel, B., Leung, J., and Dafoe, A. (2020). "The Windfall Clause: Distributing the Benefits of AI for the Common Good." *AIES '20 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Pages 327-331. Available at: https://dl.acm.org/doi/pdf/10.1145/3375627.3375842
E. [Research Report] Brundage M, Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* https://doi.org/10.17863/CAM.22520
F. [Conference Contribution] Orseau, L., & Armstrong, M. S. (2016). Safely interruptible agents. Association for Uncertainty in Artificial Intelligence. Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI), New York City, 25-29 June 2016. Available at: https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f

4. Details of the impact (indicative maximum 750 words)

Changed public understanding of risks posed by AI
Superintelligence has reached over 7,000,000 people. Over 250,000 copies of the book have been sold worldwide, and it has been translated into 31 languages. The book spawned a TED

talk with over 4,800,000 views on TED, and over 2,200,000 on YouTube as of October 2020; it was also featured in a profile in the *New Yorker*, and a BBC Hardtalk interview. [1] The book has been endorsed by numerous public figures, including Bill Gates and Elon Musk, with Musk saying:

> "Worth reading *Superintelligence.* We need to be super careful with AI." [1]

Superintelligence sparked a worldwide debate about the dangers of unsafe AI, and motivated parts of the public to change its thinking. Dozens of people publicly claimed that *Superintelligence* has changed their perception of AI risk. Reactions to the book include:

> "Bostrom has convinced me that once an AI is developed, there are many ways it can go wrong, to the detriment and possibly extermination of humanity"; or "Makes me want to change fields and work on the control problem, given that if this book is correct, it's the single most important problem humankind will ever solve." [1]

Additionally, the book **inspired the creation of art.** For example, the author Jude Mace wrote about the creation of her book *The Seed of the Violet Tree*:

> "I had time to read, explore and contemplate the possibilities and risks superintelligence presents to humanity. My interest began after reading Nick Bostrom's, *Superintelligence: Paths. Dangers, Strategies…*". [1]

Changed policy regarding future of AI

Nick Bostrom, with other FHI researchers, have informed policy decisions about AI.

The **UK House of Commons Science and Technology Committee recommended a standing Commission on AI** be established. On 24 March 2016, Dr Owen Cotton-Barratt's testimony informed the UK House of Commons Science and Technology Committee's recommendation that "a standing Commission on Artificial Intelligence be established…", which "should focus on establishing principles to govern the development and application of AI techniques, as well as advising the Government of any regulation required on limits to its progression." [2, paragraphs 59, 65/66, 73].

The **House of Lords Select Committee on Artificial Intelligence recommended more funding for AI research** at universities**.** Following testimony from Bostrom on 10 October 2017 and written evidence from FHI, the House of Lords Select Committee on Artificial Intelligence recommended a) that "the funding for PhD places in AI and machine learning be further expanded" [3, paras 163, 169], and b) that "universities and research councils providing grants and funding to AI researchers must insist that applications for such money demonstrate an awareness of the implications of the research and how it might be misused." [3, paragraphs 321, 329].

The All-Party Parliamentary Group on Artificial Intelligence recommended placing **more focus on global developments in UK AI policy**. The All-Party Parliamentary Group on Artificial Intelligence recommended on 30 October 2017 to "commission research or create a forum mapping out the AI global ecosystem and best practices from other countries and intergovernmental organisations"; and to "apply both a national and international lens when addressing AI issues", both after citing Allan Dafoe stating that "'AI is both. It is an issue on the national and international domain.' International collaboration is necessary to address many of its issues, but each government must be held responsible for setting the right policy frameworks within its own borders". [4, pp 19, 28]

Changes to institutions' policies on safe and accountable AI

Various researchers at FHI, including Bostrom, Dafoe, and Drexler, have worked closely with commercial and non-commercial partners including staff at Google Deepmind, the G30, the World Intelligence Congress in China, and the Bank of England [5]. This engagement has contributed to these institutions' emerging policies regarding ethical AI. Among other output, researchers at FHI led by Miles Brundage contributed to a report led by OpenAI, one of the largest private AI labs worldwide working on AI safety. Being produced with industry

partners (OpenAI), the document summarises leading advances in the research into AI transparency, and also **guides AI policies on verification for regulators and companies, including recommendation of stronger third-party auditing and reporting standards** on AI incidences, **and software mechanisms** like more emphasis on interpretability and privacy-preserving machine learning. [6] OpenAI stated that they will adopt several of the recommendations in their own policies [7].

Deepmind, a Google division working on advanced AI, acknowledged the value of FHI's work as well. Discussing a joint Deepmind-FHI paper on AI safety [F], the Deepmind co-author said of FHI researcher Stuart Armstrong:

> "It was a real pleasure to work with Stuart on this. … This collaboration is one of the first steps toward AI Safety research, and there's no doubt FHI and Google DeepMind will work again together to make AI safer." [8]

Development of AI safety as a focus area for philanthropic funding

Superintelligence helped to catalyse a global conversation about advanced Artificial Intelligence, which led to increased funding and interest in the field of technical AI safety, which will in turn give rise to further technological development (and resultant impact) over the long term.

For example, the Open Philanthropy Project (OPP), a U.S.-based grant-making organisation co-founded by Facebook co-founder Dustin Moskovitz, which recommended USD100,000,000 worth of grants in 2018, has committed USD114,813,767 of funding since 2015 for projects relating to potential risks from advanced artificial intelligence'. In their discussion of the reasoning underlying their decision to focus on safe AI as a key area for grant-making, the OPP recommended reading *Superintelligence* and cited Bostrom's work throughout. [9] The CEO of OPP also explicitly named *Superintelligence* as one of the reasons which convinced him of the importance of risk from advanced AI. [10] In a testimonial, OPP referred to FHI as an '*invaluable pioneer and thought partner*' for their work, that Bostrom's work had fed directly into their decision making, and that his work on astronomical waste was an important factor in OPP's early engagement with arguments about long-run future consequences, where it now concentrates half its funding. [11]

5. Sources to corroborate the impact (indicative maximum of 10 references)

1. Collection of media engagements and online testimonials for *Superintelligence*
   i. TED Talk, March 2015 (engagement statistics as at October 2020)
   ii. *New Yorker* Profile, 23 November 2015
   iii. *BBC Hardtalk* interview, 14 September 2015
   iv. Report in *The Economist*, 9 August 2014
   v. Public endorsements on social media from Elon Musk, Bill Gates, Jude Mace
   vi. Quotes from members of the public who posted on social media where they note change in thinking after reading or listening to Bostrom.
2. UK House of Commons Science and Technology Committee (2017), *Robotics and artificial intelligence: Fifth Report of Session 2016–17*, available at: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf
3. UK House of Lords Select Committee on Artificial Intelligence (2017), *AI in the UK: ready, willing and able? Report of Session 2017–19*, available at: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf
4. UK All-Party Parliamentary Group on Artificial Intelligence (2017), *International Perspective and Exemplars: a theme report based on the 7th meeting*, 30 October 2017. Available at: https://web.archive.org/web/20191221210315/http://www.appg-ai.org/wp-content/uploads/2017/12/appgai_theme_report_7_final.pdf
5. Websites and media showing a selection of industry engagements of FHI staff.
   i. Professor Bostrom's talk at *Google Deepmind* (22 September 2014)

ii. Professor Bostrom's talk at G30 (5 July 2017)
iii. Professor Bostrom's talk at the *World Intelligence Congress* (China, 23 May 2017)
iv. Professor Bostrom's seminar with staff at the Bank of England (11 April 2016)

6. Publication on recommendations for ethical safe AI developed by members of FHI (and others). Miles Brundage, …, Jade Leung, …, Carina Prunkl, …, Brian Tse, …, Allan Dafoe, …, Markus Anderljung (April 2020), "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims", available at: https://arxiv.org/pdf/2004.07213.pdf
7. OpenAI statement on implementing policies from report 'Toward Trustworthy AI', 16 April 2020. Available at: https://openai.com/blog/improving-verifiability/
8. Press Release documenting intention for Google to continue engaging with the FHI from 6 June 2016, available at: https://www.fhi.ox.ac.uk/google-deepmind-and-fhi-collaborate-to-present-research-at-uai-2016/
9. Report I from Open Philanthropy Project (August 2015) documenting the recommendation of Bostrom's book. Available at: https://www.openphilanthropy.org/research/cause-reports/ai-risk
10. Report II from Open Philanthropy Project (September 2016) *Three Key Issues I've Changed My Mind About*, available at: https://www.openphilanthropy.org/blog/three-key-issues-ive-changed-my-mind-about
11. Email statement from Programme Officer of Open Philanthropy Project, 14 October 2020.