Institution: University of Cambridge		
Unit of Assessment: UoA 30		
Title of case study: Artificial Intelligence governance, ethics, and foresight		
Period when the underpinning research was undertaken: 2015-present		
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:**
Huw Price	Academic Director of LCFI, Professor of Philosophy, Faculty of Philosophy	01.10.2011-present
Seán Ó hÉigeartaigh	Project Leader, LCFI	01.04.2015-present
Marta Halina	Project Leader, LCFI, University Lecturer (Department of History and Philosophy of Science)	08.09.2014-present
Stephen Cave	Executive Director, LCFI, Senior Research Associate, Faculty of Philosophy	04.07.2016-present
Anna Alexandrova	Project Leader, LCFI, Reader (Department of History and Philosophy of Science)	01.09.2011-present
Karina Vold	Postdoctoral Research Associate, LCFI, and Research Fellow, Faculty of Philosophy	October 2017-July 2020
Rune Nyrup	Postdoctoral Research Associate, LCFI, and Research Fellow, Department of History and Philosophy of Science	01.04.2017-present
Period when the claimed impact occurred: 2016-present		
Is this case study continued from a case study submitted in 2014? No		

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

The Leverhulme Centre for the Future of Intelligence (LCFI) was founded in 2016 to explore the short-term and long-term opportunities and challenges of artificial intelligence. Thanks to the Centre´s research and lobbying activity, governments, policymakers, and AI businesses around the world have introduced measures to improve AI governance and uphold ethical standards in the development of new AI technologies. LCFI research has led to: the inclusion of AI governance in the remit of the UK government's new Centre for Data Ethics and Innovation; contributions to national and international AI governance documents, including the report of the UK Parliamentary Select Committee on AI and strategies published by the US Government and the European Union; and changes to a number of AI company and industry policies.

2. Underpinning research (indicative maximum 500 words)

LCFI has built a new interdisciplinary community of researchers, with strong links to technologists and the policy world, and a clear practical goal: to work together to ensure that we make the best of the opportunities of artificial intelligence as it develops over coming decades. LCFI's research explores the short and long-term opportunities and challenges of a potentially epoch-making technology. The Centre was launched in 2016 with GBP10 million grant from the Leverhulme Trust. LCFI is based at the University of Cambridge, with partners at the University of Oxford, Imperial College London, and the University of California, Berkeley. Though the Centre is interdisciplinary and involves researchers from a range of departments at Cambridge, all underpinning research featured in this case study has either been led by or exclusively undertaken by philosophers at Cambridge. Most of these philosophers are based in either the Faculty of Philosophy or the Department of History and Philosophy of Science, though Sean Ó hÉigeartaigh is employed directly by LCFI and by the Centre for the Study of Existential Risk (featured in another of our impact case studies).

LCFI philosophical research in AI addresses both the dangers and the opportunities afforded by AI technology. The dangers studied by the Centre range from concerns around transparency of algorithms to the potential for AI technologies to undermine core principles of democracy. The work most relevant to LCFI's impact includes:

- Assessment of the potential for the rhetoric of AI innovation as a competitive international 'race' to lead to less responsible AI technological development. This work suggests alternative, more collaborative approaches with reduced risk. **[R1]**
- Work arguing against the common tendency in AI research to sharply distinguish short-term from long-term risks, on the grounds that neither research planning nor policy development should treat these two perspectives separately. **[R3]**
- Study of the global security risks resulting from AI or robotics-caused harm, and how international law and regulation can respond to these risks. **[R2]**
- A comprehensive, peer-reviewed report, published by the Nuffield Foundation, on the research challenges of ethical AI and related technologies. **[R4]**
- Work arguing that AI ethics in practice must move beyond the enunciation of principles towards developing processes for addressing value conflicts and trade-offs. **[R5]**
- Work arguing that the discourse of AI can perpetuate a range of biases and ideologies, such as racial prejudice. **[R6]**

This work on the risks and opportunities of AI has been complemented by the Animal AI Olympics project run by a team of LCFI researchers led by philosopher of cognitive science Marta Halina. The competition incentivises AI developers to pit their best approaches against animal intelligence to test whether cutting-edge AI technology can compete with simple animals when adapting to unexpected changes in the environment. This competition, rather than stipulating a specific task, provides a well-defined arena (launched April 2019) and a list of cognitive abilities tested in that arena. **[R7]**

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

[R1] Cave, S., **O hÉigeartaigh, S.** (2018). *An AI race for strategic advantage: Rhetoric and risks*. Association for the Advancement of Artificial Intelligence (AAAI). [Link]
The paper won Best Paper Award at the 2018 AAAI/ACM Conference on AI, Ethics and Society.
[R2] Kunz, M., **Ó hÉigeartaigh, S.** (2020). Artificial Intelligence and robotization. In R. Geiß and N. Melzer (Eds.), *Oxford handbook on the international law of global security* (16 pp.). Oxford University Press. [DOI]
[R3] Cave, S., **O hÉigeartaigh, S.** (2019). Bridging near- and long-term concerns about AI. *Nature Machine Intelligence*, 1, 5-6. [DOI]
[R4] Whittlestone, J., **Nyrup, R., Alexandrova, A.**, Dihal, K., **Cave, S.** (2019). *Ethical and societal implications of algorithms, data, and Artificial Intelligence: A roadmap for research*. The Nuffield Foundation. ISBN: 9781916021105. [Link]
[R5] Whittlestone, J., **Nyrup, R., Alexandrova, A., Cave, S.** (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 195-200. [DOI]
[R6] Cave, S., Dihal, K. (2020). The whiteness of AI. *Philosophy & Technology*, 3, 685-703. [DOI]
[R7] Crosby, M., Beyret, B., **Halina, M.** (2019). The Animal-AI Olympics. *Nature Machine Intelligence*, 1:257. [DOI]

The above outputs all meet the 2* requirement. **[R1, R3, R4, R5, R6 and R7]** have all been peer reviewed. **[R2]** was based upon research conducted as part of the Leverhulme Grant.

Research funding: LCFI is funded by grants from a variety of funders including the Templeton World Charity Foundation, Nuffield Foundation, and the Wellcome Trust. The majority of its funding comprises a GBP10 million grant awarded in 2015 by the Leverhulme Trust.

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

UK policy: ethics and governance of AI

From 2017, UK Parliament and Government undertook a programme of activity to scope and implement world-leading AI governance. LCFI played a significant role in this policy process. LCFI's biggest impact during this time was its influence on the creation and direction of a new governance body created by the UK Government – the Centre for Data Ethics and Innovation (CDEI) – the world's first national advisory body for AI.

LCFI was partly responsible for the inclusion of AI governance in the remit of the CDEI. In its report on AI, the House of Commons Science and Technology Committee agreed with Price's suggestion of a governance body **[E1, para 65]**, and recommended to the Government that 'a standing Commission on Artificial Intelligence be established' **[E1, para 73]**. The LCFI team also made the case for a new AI governance body in their Written Evidence to the House of Lords Select Committee on AI **[E2]**, and had several discussions with civil servants. LCFI researchers were consulted several times by the team within DCMS (Department for Digital, Culture, Media, and Sport) who were developing the remit for the new Centre.

Though LCFI was one of many organisations lobbying for this new form of governance, the Centre's engagement with the government team responsible for developing CDEI had a distinct influence on its existence and its remit:

> 'LCFI – and Stephen Cave especially – provided critical help to my team during its early phase of work to establish the need for a new Government body. He helped identify the nature of the ethical challenges associated with AI and heavily shaped CDEI's early programme of work. He and his team participated in a number of roundtables and contributed to a discussion paper on 'targeting' that helped us scope out the key policy questions in this space. Stephen was consistently insightful, constructive and practical – and made a real difference to the set up and design of the Centre for Data Ethics and Innovation.'
> Cora Govett, Deputy Director, Digital Regulation and Markets, UK Government's Department of Digital, Culture, Media, and Sport **[E3]**

Research Policy in the UK and China: work with the Nuffield Foundation

Over six months between 2017 and 2018, the Nuffield Foundation convened a partnership of leading UK research policy organisations and research funders to address the need for agreed ethical frameworks for the use of new technologies. This consultation resulted in the Ada Lovelace Institute (Ada), an independent research institute with a GBP5 million five-year research fund to be used to examine ethical and social issues arising from the use of data, algorithms, and AI.

LCFI played a significant role in the consultation, deploying the team's research expertise in the challenges of AI governance to shape the policy and strategy of Ada. LCFI's contribution to Ada's policy included new research commissioned by Nuffield to scope the appropriate remit for Ada's work **[R4]**. Tim Gardam, Chief Executive of the Nuffield Foundation, writes:

> 'The approach Nuffield took to developing Ada's remit was initially shaped by contributions from the Director of the LCFI at a Royal Society seminar […] Following the seminar, the Nuffield Foundation funded a research project from LCFI […] Its purpose was to inform the thinking of the Ada Lovelace Institute in its first months. LCFI's conclusions are now reflected in the Institute's mission statement and strategy.' **[E4]**

The LCFI team have also worked with the Nuffield Foundation to develop their partnerships with AI research and development organisations overseas. In November 2019, LCFI launched a new bilateral China-UK Research Centre for AI Ethics and Governance, in collaboration with Prof Yi Zeng of the Institute of Automation at the Chinese Academic of Sciences. The success of this joint venture thus far, and its potential for further research, has had a significant influence at the Nuffield Foundation, and as a result they are now investing resources into developing other similar collaborations in AI research between the UK and China **[E4]**. Ada's partnership with LCFI on global comparative research is reflected in Ada's 2019-20 Strategy **[E10, p.11]**, and the influence of LCFI is also confirmed in Gardam's testimonial statement **[E4]**.

International policy

Elsewhere, LCFI research has had considerable impact on public policy debates and strategies for AI. Over 30 national and international AI strategies have been announced in the last five years. The Centre has contributed to the drafting process for several of these. Examples include:

- In Singapore, the LCFI team has collaborated on a series of events with the Centre for Strategic Futures, based in the office of Singapore's PM. One workshop in 2018 on Risk and Artificial Intelligence helped a number of Singapore policy stakeholders navigate the new governance system introduced that year by the Singapore government (influence of LCFI in this regard confirmed by a statement from the Singapore PM's office **[E5]**). This also played a role in the development of Singapore's Model AI Governance Framework in 2019.
- In virtue of LCFI's research expertise, the UN requested that members of LCFI lead one of the four tracks at the UN's AI for Good Summit, which brought together over 30 UN agencies to discuss global AI policy. The immediate impact of the LCFI involvement in the AI for Good Summit was that LCFI research set the agenda for discussion among all UN stakeholders in AI policy debate, focussing specifically on issues around trust in AI technology **[E6]**
- LCFI research on AI applications of philosophy of mind and ethics has been cited by the Vatican's Working Group on Robotics (part of the Pontifical Academy for Life) in a report on the Vatican's event in 2020 to mark the new Rome Call for AI Ethics, a statement regarding the need for human-centric AI ethics **[E7, fn.4 and 21]**.

Impact on the AI industry

LCFI research has also had an impact on company policy within the AI industry, both within the UK and internationally. In February 2018, LCFI hosted experts in AI and security at a workshop held in Oxford. Participants included Google, Microsoft, Deepmind, and OpenAI, a research laboratory based in San Francisco. One outcome of this workshop, and other interactions with OpenAI, was influence on their April 2018 Charter. The Charter references the importance of avoiding competitive development races, echoing concerns raised by LCFI researchers at the workshop and in published outputs **[R1; E11]**. The Centre has also contributed to the work of Digital Catapult, a UK government innovation agency for the digital industry, helping computing start-up companies. In September 2018, Digital Catapult launched a new AI Ethics Framework for AI start-ups, partly influenced by contributions from LCFI researchers **[E12]**.

In addition to impact on ethics and governance in company policy, LCFI researchers have worked with AI company GoodAI to run an innovation competition to inspire and influence AI developers and prompt responsible AI research and development. GoodAI has chosen the Animal-AI Olympics **[R7]** as its General AI Challenge for 2019 **[E8]**. Olga Afanasjeva, COO at GoodAI, told the team:

> 'Before the Animal-AI Olympics we were lacking a comprehensive curriculum for AI which would be based on state-of-the-art techniques in measuring animal intelligence…The project also provides a publicly available benchmark for adaptive behavior in a realistic setting. We expect this new opportunity for comparing results between different scientific teams to be a driver for research in the areas of transfer learning, curriculum learning and self-supervised learning. The Animal-AI Olympics environment is now added to our research roadmap and will play an instrumental role in our work in the future.' **[E8]**

Participant developers have also confirmed the significance of the Olympics for their work. Dymtro Bobrenko, Machine Learning Engineer for Samsung Electronics, told the team: 'Animal-AI Olympics gave us a unique opportunity to benchmark our Motion Planning algorithms by competing with the world's best solutions in current domain. Currently, there is no standardized benchmark for Motion Planning and AI in general, and without Animal-AI competition it would be difficult to evaluate our algorithms objectively.' **[E9]**

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

[E1] House of Commons: Science and Technology Select Committee. (2016). *Robotics and artificial intelligence. Fifth Report of Session 2016–17*. [Link]
[E2] Written evidence to House of Lords and Select Committee on Artificial Intelligence Report "AI In the UK: ready, willing and able?" (2018). pp.173-4.
[E3] Testimonial from Deputy Director, Digital Regulation and Markets, DCMS
[E4] Testimonial from Chief Executive at the Nuffield Foundation
[E5] Testimonial from the Singapore PM's Office
[E6] Webpage confirming LCFI team members leading one of the four UN AI tracks. See Track 4: Trust in AI (authors from this case study highlighted). [Link]
[E7] Sinibaldi et al. (2020). Contributions from the Catholic Church to ethical reflections in the digital era, Nature Machine Intelligence, 2, 242–244. [DOI]
Comment piece written by researchers at the Vatican's Working Group on Robotics, part of the Pontifical Academy for Life
[E8] Testimonial from GoodAI
[E9] Testimonial from Samsung Electronics Machine Learning Engineer
[E10] Ada Lovelace Institute. Our Strategy 2019-2020. [Link]
[E11] Webpage: OpenAI charter. [Link]
[E12] Digital Catapult Ethics Framework [Link]