Futures xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Accompanying technology development in the Human Brain Project: From foresight to ethics management

Christine Aicardi[a,*], B. Tyr Fothergill[b], Stephen Rainey[c], Bernd Carsten Stahl[b], Emma Harris[b]

[a] Department of Global Health & Social Medicine, King's College London, United Kingdom
[b] Centre for Computing and Social Responsibility, De Montfort University, United Kingdom
[c] The Oxford Uehiro Centre for Practical Ethics, University of Oxford, United Kingdom

ARTICLE INFO

Keywords:
Responsible research and innovation
Human Brain Project
Foresight
Ethics management
Artificial intelligence

ABSTRACT

This paper addresses the question of managing the existential risk potential of general Artificial Intelligence (AI), as well as the more near-term yet hazardous and disruptive implications of specialised AI, from the perspective of a particular research project that could make a significant contribution to the development of Artificial Intelligence (AI): the Human Brain Project (HBP), a ten-year Future and Emerging Technologies Flagship of the European Commission. The HBP aims to create a digital research infrastructure for brain science, cognitive neuroscience, and brain-inspired computing. This paper builds on work undertaken in the HBP's Ethics and Society subproject (SP12). Collaborators from two activities in SP12, Foresight and Researcher Awareness on the one hand, and Ethics Management on the other, use the case of machine intelligence to illustrate key aspects of the dynamic processes through which questions of ethics and society, including existential risks, are approached in the organisational context of the HBP. The overall aim of the paper is to provide practice-based evidence, enriched by self-reflexive assessment of the approach used and its limitations, for guiding policy makers and communities who are, and will be, engaging with such questions.

1. Introduction

Existential risks, 'x-risks' for short, are commonly understood as hypothetical future events that could cause the extinction of humanity or drastically alter its continued existence. The existential risks associated with technological developments have attracted much attention in the recent past, with the creation of dedicated institutions such as the Future of Life Institute (founded in 2014),[1] or the Centre for the Study of Existential Risk at the University of Cambridge (founded in 2012),[2] a concern of which is to bring together "the 'x-risk ecosystem' – a thriving community of researchers and others, inside and outside academia, united by a common interest in potential serious hazards of powerful and beneficial new technologies […] to ask ourselves where our efforts should best be directed, over the rest of the decade and beyond."[3]

This paper aims to contribute to the debate by focusing on the risks (not all existential yet no less serious) posed by one such

* Corresponding author.
E-mail addresses: christine.aicardi@kcl.ac.uk (C. Aicardi), tyr.fothergill@dmu.ac.uk (B.T. Fothergill), stephen.rainey@philosophy.ox.ac.uk (S. Rainey), bstahl@dmu.ac.uk (B.C. Stahl), emma.harris@dmu.ac.uk (E. Harris).

[1] https://futureoflife.org/, consulted 10/01/2018.
[2] https://www.cser.ac.uk/, consulted 10/01/2018.
[3] http://www.crassh.cam.ac.uk/events/27021, consulted 19/12/2017.

https://doi.org/10.1016/j.futures.2018.01.005
Received 13 April 2017; Received in revised form 3 January 2018; Accepted 17 January 2018

technology, artificial intelligence (AI). The topic of machine intelligence as a potential threat to humanity is not new. It has long been a theme in popular culture, the archetypal mad scientists Faust and Frankenstein established the powerful trope of pessimism about scientific endeavours and a fear of their results (Weingart, 2010, p. 339). Modern cinema has tended to reinforce these concerns, particularly when depicting machine intelligence. Films such as *Terminator* (dir. Cameron, 1984), *The Matrix* (dir. Wachowskis, 1999) and *Transcendence* (dir. Pfister, 2014) depict machine intelligence as dangerous and destructive, though it should be noted that several recent films and TV series have been more ambiguous in this regard and that video games such as *Mass Effect: Andromeda* (dev. BioWare, 2017) are optimistic in their depictions.

These cultural trends may help to explain why the Special Eurobarometer 382: Public Attitudes towards European Commission (2012) found such negative attitudes to AI and robotics in 'human' roles. The survey found that a large majority of respondents were sceptical or fearful of machine intelligence becoming part of their personal, as opposed to professional, lives. '[T]here is widespread agreement that robots should be banned in the care of children, the elderly or the disabled (60%) with large minorities also wanting a ban when it comes to other "human" areas such as education (34%), healthcare (27%) and leisure (20%)' (2012: 4).

More recently, the topic has attracted a high level of attention, as indicated by the UK Parliament's Science and Technology Select Committee's report on "robotics and artificial intelligence" (House of Commons Science and Technology Committee, 2016) which mirrors reports on the same topic from the US (Executive Office of the President, 2016a,b) and the European Parliament (Committee on Legal Affairs, 2017). This heightened attention by policymakers reflects a growing awareness that the confluence of artificial intelligence techniques, big data, high processing power at low energy cost, and the increasing spread of information and communication technologies (ICTs) has arrived at the point where it can plausibly be said to have potentially significant impact on people's lives.

This growing awareness of the increasing power of AI does not by itself imply that these technologies pose a particular risk, even less that they pose an existential risk in the sense that they threaten the very survival of humanity or at least of our current way of life. They are nevertheless a good starting point to ask whether such risks may materialise and how they could be addressed.

This paper addresses that question from the perspective of a research project with the potential to make a significant contribution to the development of AI. The Human Brain Project (HBP, www.humanbrainproject.eu), a ten-year Future and Emerging Technologies Flagship initiative of the European Commission, has the overall aim to create an ICT-based scientific research infrastructure for brain research, cognitive neuroscience, and brain-inspired computing. To this end, it brings together a number of activities, including animal, human, cognitive, and theoretical neuroscience as well as platform development in the fields of neuroinformatics, high performance analytics and computing, medical informatics, neuromorphic computing, and neurorobotics. This combination of activities offers the possibility of ground-breaking insights that can substantially change or accelerate the development of artificial intelligence. The exact capabilities of these new technologies are still difficult to assess, but in seeking to capitalise on our understanding of animal and human brains, we have high expectations regarding their impact. The flipside of these hopes for the development of novel technologies is that they may constitute risks that are difficult or even impossible to evaluate.

From early on in its development, the HBP has been aware of these and other social, ethical and philosophical concerns, and has dedicated a set of activities to such questions. These are organised around the principles of Responsible Research and Innovation (RRI). RRI, in the interpretation adopted by the UK Engineering and Physical Research Council (EPSRC) through its AREA framework (Anticipate, Reflect, Engage, Act),[4] suggests that research needs to include anticipation of possible future consequences, reflection on the rationale and justification of research, engagement with various stakeholders, and translation of these activities into action. It is hoped that incorporating these principles in all aspects of the research and innovation process will make it more socially responsible. The HBP has implemented these principles through four work packages in Subproject SP12, Ethics and Society, which cover foresight and researcher awareness, conceptual and philosophical reflection, public engagement, and ethics management. Although existential risks are unprecedented and thus particularly difficult to identify, these activities will hopefully detect if the HBP starts raising existential risks, and in any case recognize other serious risks, before recommending suitable ways of addressing them, and developing appropriate action plans.[5]

This present paper builds on work undertaken in the HBP's Ethics and Society subproject (SP12). Collaborators from two tasks and a work package in SP12, Foresight, Researcher Awareness, and Ethics Management use the case of machine intelligence to illustrate key aspects of the dynamic process through which questions of ethics and society, including existential risks, are approached in the HBP organisation. The overall aim of the paper is to provide practice-based evidence, enriched by the self-reflexive assessment of the approach used and its limitations, for guiding policy makers and communities who are, and will be, engaging with such questions.

The foundational work was initiated in the Ramp-Up Phase of the HBP (between October 2013 and March 2016) and continues into the 1st tranche of the HBP Operational Phase (SGA1, between April 2016 and March 2018), around the potential contribution that the Project could make to future computing and robotics, machine intelligence in particular.

Firstly, this paper briefly presents the conclusions of the foresight work that was conducted during the Ramp-Up Phase of the HBP. It then details how the resulting recommendations are being developed for action by the Ethics Management and the Researcher Awareness teams. It thereby demonstrates how researcher awareness and ethics management can evaluate and act on the issues initially raised in foresight work, and take them back to the researchers and other members of the HBP in order to increase their capacity to reflect on ethical, social, and regulatory issues, thus helping close the loop between anticipation and action in the AREA

[4] https://www.epsrc.ac.uk/research/framework/area/, consulted 19/12/2017.

[5] For detailed perspectives on the role and activities of the Ethics and Society Subproject in the HBP, see (Aicardi et al., 2017; Evers, 2017; Rainey, Stahl, Shaw, & Reinsborough, 2017).