

# Joint written evidence submitted by Future of Humanity Institute, Centre for the Study of Existential Risk, Global Priorities Project, and Future of Life Institute (ROB0052)

The Future of Humanity Institute, Centre for the Study of Existential Risk, Global Priorities Project, and Future of Life Institute are multi-disciplinary research institutions that focus on future risks to humanity, including those posed by artificial intelligence. Several individual researchers associated with these centres have submitted additional evidence in a personal capacity, expanding on areas where they have specialist expertise.

Artificial intelligence (AI) will likely act across a wide range of political, social, and economic dimensions, making it more appropriate to think of AI as a collection of issues than as a single issue. Given possibilities for path-dependence or lock-in, governments should ensure that short- and medium-term AI policies remain compatible with longer-term objectives. We identify five distinct policy areas related to AI.

1. **Data & privacy:** How should we protect personal privacy as machine learning and big data enable companies and governments to gain unprecedented access to people's lives?
2. **Autonomous systems & liability:** Given the increasing reliance on intelligent systems to make decisions, who is responsible in circumstances when the consequences are harmful?
3. **Automation & unemployment:** When growing sections of the workforce lack the skills to compete against automated systems in the labor market, how do we maintain the standards of welfare and stability necessary for a functioning society? How do other nations with less skilled workforces maintain stability? What are the effects on the UK if they cannot?
4. **Military, security, and geopolitical coordination:** With advanced (lethal and non-lethal) automated weapons, submersible long-distance drones, and vulnerability to cyber warfare, will the relatively peaceful period following WWII be maintainable? Will these technologies undermine the doctrine of mutually assured destruction? How significant is the risk that these technologies will be used in unintended ways by individuals and nations with very different interests than the UK? What avenues are available for international coordination, cooperation, and stability to prevent this?
5. **Implications of superintelligent AI:** How can we coordinate to ensure the safe development of superintelligent AI, both in terms of it being controllable, and ensuring that it is used in a way beneficial to the interests of all humanity? The Future of Humanity Institute's Director, Nick Bostrom, discusses these issues in *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

Exactly how these different issues should be addressed is unclear. The first two could be an appropriate subject for a Warnock-style committee. The third and fourth issues may be worth studying within Government, for example, in the case of the third issue, through BIS. On the last issue, we feel there is not yet enough expert consensus to address the topic at a governmental level. Additionally, with high risks of precedential lock-in and path dependencies, attempting to set policy prematurely might be irresponsible. However, we believe that it might be worth returning to this issue at a governmental level in 5-10 years when it comes into clearer focus within the research community.

We additionally recommend that policymakers heed the following general principles when addressing short- and medium-term issues to ensure that those actions are compatible with longer-term considerations.

Joint written evidence submitted by Future of Humanity Institute, Centre for the Study of Existential Risk, Global Priorities Project, and Future of Life Institute (ROB0052)

1. Government should be careful not to take actions which are based on an implicit premise about AI having a clear upper limit in future capability.
2. Groups investigating these issues may benefit from focusing on broad principles and not narrow recommendations. These should be flexible in application to new developments and increased capability.
3. Government should keep a broadly global view in its approach on AI. It is possible that AI advances will be widely destabilizing, and that some nations will not adapt well. Both the UK, and the rest of the world, will benefit if the UK can lay out broad, cooperative, and replicable frameworks which favour stability, economic well-being, and coordination.

*April 2016*