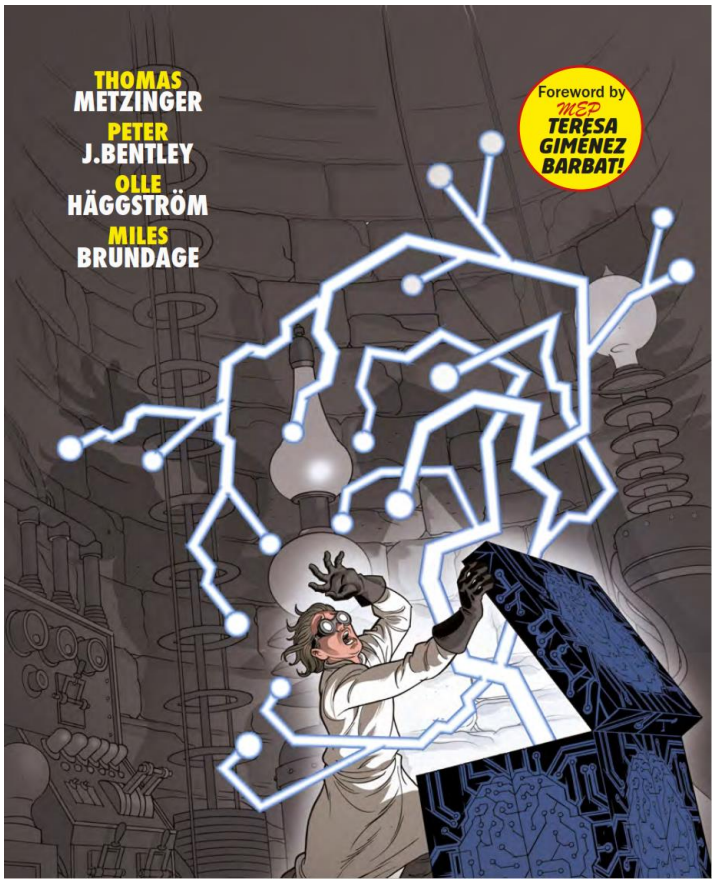
European Parliament

Should we fear artificial intelligence?



IN-DEPTH ANALYSIS

Science and Technology Options Assessment

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

March 2018 - PE 614.547

6. Towards a Global Artificial Intelligence Charter

Thomas Metzinger

Introduction

It is now time to move the ongoing public debate on artificial intelligence (AI) into the political institutions themselves. Many experts believe that we are confronted with an inflection point in history during the next decade, and that there is a closing time window regarding the applied ethics of AI. Political institutions must therefore produce *and* implement a minimal, but sufficient set of ethical and legal constraints for the beneficial use and future development of AI. They must also create a rational, evidence-based process of critical discussion aimed at continuously updating, improving and revising this first set of normative constraints. Given the current situation, the default outcome is that the values guiding AI development will be set by a very small number of human beings, by large private corporations and military institutions. Therefore, one goal is to proactively integrate as many perspectives as possible – and in a timely manner.

Many different initiatives have already sprung up world-wide and are actively investigating recent advances in AI in relation to issues concerning applied ethics, its legal aspects, future sociocultural implications, existential risks and policy-making.[1] There exists a heated public debate, and some may even gain the impression that major political institutions like the EU are not able to react in an adequate speed to new technological risks and to rising concern in the general public. We should therefore increase the agility, efficiency and systematicity of current political efforts to implement rules by developing a more formal and institutionalised democratic process, and perhaps even new models of governance.

To begin a more systematic and structured process, I will present a concise and non-exclusive list of the five most important problem domains, each with practical recommendations. The first problem domain to be examined is the one which, in my view, is constituted by those issues having the smallest chances to be solved. It should therefore be approached in a multi-layered process, beginning in the European Union (EU) itself.

The "race-to-the-bottom" problem

We need to develop and implement world-wide safety standards for AI research. A *Global* Charter for AI is necessary, because such safety standards can only be effective if they involve a binding commitment to certain rules by *all* countries participating and investing in the relevant type of research and development. Given the current competitive economic and military context, the safety of AI research will very likely be reduced in favour of more rapid progress and reduced cost, namely by moving it to countries with low safety standards and low political transparency (an obvious, strong analogy is the problem of tax evasion by corporations and trusts). If international cooperation and coordination succeeds, then a "race to bottom" in safety standards (through the relocation of scientific and industrial AI research) could in principle be avoided. However, the currently given landscape of incentives makes this a highly unlikely outcome.

---

[1] For an overview of existing initiatives, see for example Baum 2017 and Boddington 2017, p. 3p. I have refrained from providing full documentation here, but helpful entry points into the literature are Mannino et al. 2015, Stone et al. 2016, IEEE 2017, Bostrom, Dafoe & Flynn 2017, Madary & Metzinger 2016 (for VR).

Recommendation 1

The EU should immediately develop a European AI Charter.

Recommendation 2

In parallel, the EU should initiate a political process leading the development of an Global AI Charter.

Recommendation 3

The EU should invest resources into systematically strengthening international cooperation and coordination. Strategic mistrust should be minimised, commonalities can be defined via maximally negative scenarios.

The second problem domain to be examined, is arguably constituted by the most urgent set of issues, and these also have a rather small chance to be solved to a sufficient degree.

Prevention of an AI arms race

It is in the interest of the citizens of EU that an AI arms race, for example between China and the US, is prevented at a very early stage. Again, it may well be too late for this, and obviously European influence is limited, but research into and development of offensive autonomous weapons should be banned and not be funded on EU territory. Autonomous weapons select and engage targets without human intervention, they will act on ever shorter time- and reaction-scales, which in turn will make it rational to transfer more and more human autonomy into these systems themselves. They may therefore create military contexts in which it is rational to relinquish human control almost entirely. In this problem domain, the degree of complexity is even higher than in preventing the development and proliferation nuclear weapons, for example, because most of the relevant research does not take place in public universities. In addition, if humanity forces itself into an arms race on this new technological level, the historical process of an arms race *itself* may become autonomous and resist political interventions.

Recommendation 4

The EU should ban *all* research on offensive autonomous weapons on its territory, and seek international agreements.

Recommendation 5

For purely defensive military applications, the EU should fund research into the maximal degree of autonomy for intelligent systems that appears to be acceptable from an ethical and legal perspective.

Recommendation 6

On an international level, the EU should start a major initiative to prevent the emergence of an AI arms race, using all diplomatic and political instruments available.

The third problem domain to be examined is the one for which the predictive horizon is probably still quite distant, but where epistemic uncertainty is high and potential damage could be extremely large.

A moratorium on synthetic phenomenology

It is important that all politicians understand the difference between artificial intelligence and artificial consciousness. The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. “Synthetic phenomenology” (SP; a term coined in analogy to “synthetic biology”) refers to the possibility of creating not only general intelligence, but also consciousness or subjective experiences on advanced artificial systems. Future

artificial subjects of experience have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing negative states like suffering.[2] One potential risk is to dramatically increase the overall amount of in suffering the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.

Recommendation 7

The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements.[3]

Recommendation 8

Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience and computer science). Specific relevant topics are evidence-based conceptual, neurobiological and computational models of conscious experience, self-awareness and suffering.

Recommendation 9

On the level of foundational research there is a need to promote, fund and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness and subjectively experienced suffering.

The next general problem domain to be examined is the one which is the most complex one and which likely contains the largest number of unexpected problems and "unknown unknowns".

Dangers to social cohesion

Advanced AI technology will clearly provide many possibilities to optimise the political process itself, including novel opportunities for rational, value-based social engineering and more efficient, evidence-based forms of governance. On the other hand, it is not only plausible to assume that there are many new, at present unknown, risks and dangers potentially undermining the process of keeping our societies coherent; it is also rational to assume the existence of a larger number of "unknown unknowns", of AI-related risks that we will only discover by accident and at a late stage. Therefore, the EU should allocate *separate resources* to prepare for situations, in which such unexpected "unknown unknowns" are suddenly discovered.

Many experts believe that the most proximal and well-defined risk is massive unemployment through automatisation. The implementation of AI technology by financially potent stakeholders may therefore lead to a steeper income gradient, increased inequality, and dangerous patterns of social stratification. Concrete risks are extensive wage cuts, a collapse of income tax, plus an overload of social security systems. But AI poses many other risks for social cohesion, for example by privately owned and autonomously controlled social media aimed at harvesting human attention, and "packaging" it for further use by customers, or in "engineering" the formation of political will via Big Nudging strategies and AI-controlled choice architectures, which are not transparent to the individual citizens whose

[2] See Metzinger 2013, 2017.

[3] This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness. For recent examples see Dehaene, Lau & Kouider 2017, Graziano 2017, Kanai 2017.

behaviour is controlled in this way. Future AI technology will be extremely good at modelling and predictively controlling human behaviour – for example by positive reinforcement and indirect suggestions, making compliance with certain norms or the "spontaneous" appearance of "motives" and decision appear as entirely unforced. In combination with Big Nudging and predictive user control, intelligent surveillance technology could also increase global risks by *locally* helping to stabilise authoritarian regimes in an efficient manner. Again, very likely, most of these risks to social cohesion are still unknown at present, and we may only discover them by accident. Policy-makers must also understand that any technology that can purposefully optimise the intelligibility of its own action to human users can in principle also optimise for *deception*. Great care must therefore be taken to avoid accidental or even intended specification of the reward function of any AI in a way that might indirectly damage the common good.

AI technology currently is a private good. It is the obligation of democratic political institutions to turn large portions of it into a well-protected *common* good, something that belongs to all of humanity. In the tragedy of the commons, everyone can often see what is coming, but if mechanisms for effectively counteracting the tragedy aren't in existence it will unfold, for example in decentralised situations. The EU should proactively develop such mechanisms.

Recommendation 10

Within the EU, AI-related productivity gains must be distributed in a socially just manner. Obviously, past practice and global trends clearly point into the opposite direction: We have (almost) never done this in the past, and existing financial incentives directly counteract this recommendation.

Recommendation 11

The EU should carefully research the potential for an unconditional basic income or a negative income tax on its territory.

Recommendation 12

Research programs are needed about the feasibility of accurately timed retraining initiatives for threatened population strata towards creative skills and social skills.

The next problem domain is difficult to tackle, because most of the cutting-edge research in AI has already moved out of publicly funded universities and research institutions. It is in the hands of private corporations, and therefore systematically non-transparent.

Research ethics

One of the most difficult theoretical problems lies in defining the conditions under which it would be rational to relinquish specific AI research pathways altogether (for instance those involving the emergence of synthetic phenomenology, or an explosive evolution of autonomously self-optimising systems not reliably aligned with human values). What would be concrete, minimal scenarios justifying a moratorium on certain branches of research? How will democratic institutions deal with deliberately unethical actors in a situation where collective decision-making is unrealistic and graded, non-global forms of *ad hoc* cooperation have to be created? Similar issues have already occurred in so called "gain-of-function research" involving experimentation aiming at an increase in the transmissibility and/or virulence of pathogens, such as certain highly pathogenic H5N1 influenza virus strains, smallpox or anthrax. Here, influenza researchers laudably imposed a voluntary and temporary moratorium on themselves. In principle, this could be possible in the AI research community as well. Therefore, the EU should always complement its AI charter with a concrete code of ethical conduct for researchers working in funded projects.

However, the deeper goal would be to develop a more comprehensive *culture of moral sensitivity* within the relevant research communities themselves. A rational, evidence-based identification and minimisation of risks (also those pertaining to a more distant future) ought to be a part of research itself and scientists should cultivate a proactive attitude, especially if they are the first to become aware of novel types of risks through their own work. Communication with the public, if needed, should be self-initiated, an act of taking control and acting in advance of a future situation, rather than just reacting to criticism by non-experts with some set of pre-existing, formal rules. As Madary and Metzinger (2016, p. 12) write in their ethical code of conduct including recommendations for good scientific practice in virtual reality: "Scientists must understand that following a code of ethics is not the same as *being* ethical. A domain-specific ethics code, however consistent, developed and fine-grained future versions of it may be, can never function as a substitute for ethical reasoning itself."

Recommendation 13

Any AI Global Charter, or its European precursor, should always be complemented by a concrete Code of Ethical Conduct guiding researchers in their practical day-to-day work.

Recommendation 14

A new generation of applied ethicists specialised on problems of AI technology, autonomous systems and related fields has to be trained. The EU should systematically and immediately invest in developing the future expertise needed within the relevant political institutions, and it should do so aiming at an above-average, especially high level of academic excellence and professionality.

Meta-governance and the pacing gap

As briefly pointed out in the introductory paragraph, the accelerating development of AI has perhaps become the *paradigmatic* example of an extreme mismatch between existing governmental approaches and what would be needed in terms of optimising the risk/benefit ratio in a timely fashion. It has become a paradigmatic example of time pressure, in terms of rational and evidence-based identification, assessment and management of emerging risks, the creation of ethical guidelines, and implementing an enforceable set of legal rules. There is a "pacing problem": Existing governance structures simply are not able to respond to the challenge fast enough; political oversight has already fallen far behind technological evolution.[4]

I am not drawing attention to the current situation because I want to strike an alarmist tone or to end on a dystopian, pessimistic note. Rather, my point is that the adaptation of governance structures *themselves* is part of the problem landscape: In order to close or at least minimise the pacing gap we have to invest resources into changing the structure of governance approaches themselves. "Meta-governance" means just this: a governance *of* governance in facing the risks and potential benefits of an explosive growth in specific sectors of technological development. For example, Wendell Wallach has pointed out that the effective oversight of emerging technologies requires some combination of both hard regulations enforced by government agencies and expanded soft governance mechanisms.[5]

[4] Gary Marchant (2011) puts the general point very clearly in the abstract of a recent book chapter: "*Emerging technologies are developing at an ever-accelerating pace, whereas legal mechanisms for potential oversight are, if anything, slowing down. Legislation is often gridlocked, regulation is frequently ossified, and judicial proceedings are sometimes described as proceeding at a glacial pace. There are two consequences of this mismatch between the speeds of technology and law. First, some problems are overseen by regulatory frameworks that are increasingly obsolete and outdated. Second, other problems lack any meaningful oversight altogether. To address this growing gap between law and regulation, new legal tools, approaches and mechanisms will be needed. Business as usual will not suffice*".

[5] See Wallach 2015 (Chapter 14), p. 250.

Marchant and Wallach have therefore proposed so-called "Governance Coordination Committees" (GCCs), a new type of institution providing a mechanism to coordinate and synchronise what they aptly describe as an "explosion of governance strategies, actions, proposals, and institutions"[6] with existing work in established political institutions. A GCC for AI could act as an "issue manager" for one specific, rapidly emerging technology, as an information clearinghouse, an early warning system, an instrument of analysis and monitoring, an international best-practice evaluator, and as an independent and trusted "go-to" source for ethicists, media, scientists and interested stakeholders. As Marchant and Wallach write: "*The influence of a GCC in meeting the critical need for a central coordinating entity will depend on its ability to establish itself as an honest broker that is respected by all relevant stakeholders*".[7]

Many other strategies and governance approaches are of course conceivable. This is not the place to discuss details. Here, the general point is simply that we can only meet the challenge posed by the rapid development in AI and autonomous systems if we put the question of meta-governance on top of our agenda right from the very beginning.

Recommendation 15

The EU should invest in researching and developing new governance structures that dramatically increase the speed by which established political institutions can respond to problems and actually enforce new regulations.

Conclusion

I have proposed that the EU immediately begins working towards the development of a Global AI Charter, in a multi-layered process starting with an AI Charter for the European Union itself. To briefly illustrate some of the core issues from my own perspective as a philosopher, I have identified five major thematic domains and provided fifteen general recommendations for critical discussion. Obviously, this contribution was not meant as an exclusive or exhaustive list of the relevant issues. On the contrary: At its core, the applied ethics of AI is not a field for grand theories or ideological debates at all, but mostly a problem of sober, rational risk management involving different predictive horizons under great uncertainty. However, an important part of the problem is that we cannot rely on intuitions, because we must satisfy counterintuitive rationality constraints.

Let me end by quoting from a recent policy paper titled *Artificial Intelligence: Opportunities and Risks*, published by the Effective Altruism Foundation in Berlin, Germany:

In decision situations where the stakes are very high, the following principles are of crucial importance:

1. Expensive precautions can be worth the cost even for low-probability risks, provided there is enough to win/lose thereby.
2. When there is little consensus in an area amongst experts, epistemic modesty is advisable. That is, one should not have too much confidence in the accuracy of one's own opinion either way.[8]

[6] This quote is taken from an unpublished, preliminary draft entitled „An agile ethical/legal model for the international and national governance of AI and robotics"; see also Marchant & Wallach 2015.

[7] Marchant & Wallach 2015, p. 47.

[8] Cf. Mannino et al. 2015.

References

Adriano, Mannino; Althaus, David; Erhardt, Jonathan; Gloor, Lukas; Hutter, Adrian; Metzinger, Thomas (2015): Artificial Intelligence. Opportunities and Risks. In: *Policy Papers of the Effective Altruism Foundation* (2), S. 1–16. https://ea-foundation.org/files/ai-opportunities-and-risks.pdf.

Baum, Seth (2017): A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. https://ssrn.com/abstract=3070741.

Boddington, Paula (2017): Towards a Code of Ethics for Artificial Intelligence. Cham: Springer International Publishing (Artificial Intelligence: Foundations, Theory, and Algorithms).

Bostrom, Nick; Dafoe, Allan; Flynn, Carrick (2017): Policy Desiderata in the Development of Machine Superintelligence. working Paper, Oxford University. http://www.nickbostrom.com/papers/aipolicy.pdf.

Dehaene, Stanislas; Lau, Hakwan; Kouider, Sid (2017): What is consciousness, and could machines have it? In: *Science (New York, N.Y.)* 358 (6362), S. 486–492. DOI: 10.1126/science.aan8871.

Graziano, Michael S. A. (2017): The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness. In: *Frontiers in Robotics and AI* 4, S. 61. DOI: 10.3389/frobt.2017.00060.

Madary, Michael; Metzinger, Thomas K. (2016): Real virtuality. A code of ethical conduct. recommendations for good scientific practice and the consumers of VR-technology. In: *Frontiers in Robotics and AI* 3, S. 3. http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full

Marchant, Gary E. (2011): The growing gap between emerging technologies and the law. In Marchant, Gary E.; Allenby, Braden R.; Herkert, Joseph R. (Hg.): The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: Springer, S. 19–33.

Marchant, Gary E.; Wallach, Wendell (2015): Coordinating technology governance. In: *Issues in Science and Technology* 31 (4), S. 43.

Metzinger, Thomas (2013): Two principles for robot ethics. In: In Hilgendorf, Eric; Günther, Jan-Philipp (Hg.) (2013): Robotik und Gesetzgebung: BadenBaden, Nomos S. 247–286. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf

Metzinger, Thomas (2017): Suffering. In: Kurt Almqvist und Anders Haag (Hg.): The Return of Consciousness. Stockholm: Axel and Margaret Ax:son Johnson Foundation, S. 237–262. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_Suffering_2017.pdf

Kanai, Ryota (2017): We Need Conscious Robots. How introspection and imagination make robots better. In: *Nautilus* (47). http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Stone, Peter; et al. (2016): Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford, CA: Stanford University. https://ai100.stanford.edu/2016-report.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017): Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html.

Wallach, W. (2015): A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control. New York: Basic Books.