Towards a Global Artificial Intelligence Charter

Thomas Metzinger*

I. INTRODUCTION

The time has come to move the ongoing public debate on Artificial Intelligence (AI) into our political institutions. Many experts believe that during the next decade we will be confronted with an inflection point in history and that there is a closing window of opportunity for working out 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 acting within large private corporations and military institutions. Therefore, one goal is to proactively integrate as many perspectives as possible – and in a timely manner. Many initiatives have already sprung up worldwide and are actively investigating recent advances in AI in relation to issues concerning applied ethics, including its legal aspects, future sociocultural implications, existential risks, and policymaking.¹ Public debate is heated, and some may even have the impression that major

* This is an updated and considerably expanded version of a chapter that goes back to a lecture I gave on 19 October 2017 at the European Parliament in Brussels (Belgium). Cf. (2018), Towards a Global Artificial Intelligence Charter. In European Parliament (ed), Should we fear artificial intelligence? PE 614.547. www.philosophie.fbo5.uni-mainz.de/ files/2018/03/Metzinger_2018_Global_Artificial_Intelligence_Charter_PE_614.547.pdf.

¹ For an overview of existing initiatives, I recommend T Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2020) 30 Minds & Machines 99 https://doi.org/10.1007/s11023-020-09517-8; and the AI Ethics Guidelines Global Inventory created by Algorithm Watch, at https://inventory.algorithmwatch.org/. Other helpful overviews are S Baum, 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy' (2017) Global Catastrophic Risk Institute Working Paper 17-1 https://ssrn.com/abstract=3070741; P Boddington, Towards a Code of Ethics for Artificial Intelligence (2017) 3. I have refrained from providing full documentation here, but useful entry points into the literature are A Mannino and others, 'Artificial Intelligence. Opportunities and Risks' (2015) 2 Policy Papers of the Effective Altruism Foundation https://ea-foundation.org/files/ai-opportunities-and-risks.pdf (hereafter Mannino et al., 'Opportunities and Risks'); P Stone and others, 'Artificial Intelligence and Life in 2030' (2016) One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel https://aioo.stanford.edu/2016-report; The IEEE Clobal Initiative on Ethics of Autonomous and Intelligent Systems, 'Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems' (IEEE, 2017) http://standards.ieee.org/develop/indconn/ ec/auto_sys_form.html; N Bostrom, A Dafoe, and C Flynn, 'Policy Desiderata in the Development of Machine Superintelligence' (2017) Oxford University Working Paper www.nickbostrom.com/papers/aipolicy.pdf; M Madary and T Metzinger, 'Real Virtuality. A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology' (2016) 3 Frontiers in Robotics and AI 3 http://journal.frontiersin.org/article/10.3389/ frobt.2016.00003/full.

political institutions like the European Union (EU) are unable to react with adequate speed to new technological risks and to rising concern amongst 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 initiate a more systematic and structured process, I will present a concise and nonexclusive list of the five most important problem domains, each with practical recommendations. The first problem domain to be examined is the one that, in my view, is made up of those issues that have the smallest chance of being solved. It should, therefore, be approached in a multilayered process, beginning in the EU itself.

II. THE 'RACE-TO-THE-BOTTOM' PROBLEM

We need to develop and implement worldwide safety standards for AI research. A Global Charter for AI is necessary, because such safety standards can be effective only 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 and strong analogy is the problem of tax evasion by corporations and trusts). If international cooperation and coordination succeeded, then a 'race to the bottom' in safety standards (through the relocation of scientific and industrial AI research) could, in principle, be avoided. However, the current landscape of incentives makes this a highly unlikely outcome. Non-democratic political actors, financiers, and industrial lobbyists will almost certainly prevent any more serious globalised approach to AI ethics.² I think that, for most of the goals I will sketch below, it would not be intellectually honest to assume that they can actually be realised, at least not in any realistic time frame and with the necessary speed (this is particularly true of Recommendations 2, 4, 6, 7, 10, 12, and 14). Nevertheless, it may be helpful to formulate a general set of desiderata to help structure future debates.

Recommendation 1

The EU should immediately develop a European AI Charter.

Recommendation 2

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

Recommendation 3

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

² T Metzinger, 'Ethics Washing Made in Europe' *Tagesspiegel* (8 April 2019) www.tagesspiegel.de/politik/eu-guidelinesethics-washing-made-in-europe/24195496.html. The second problem domain to be examined is arguably constituted by the most urgent set of issues, and these also have a fairly small chance of being adequately resolved.

III. PREVENTION OF AN AI ARMS RACE

It is in the interests of the citizens of the EU that an AI arms race, for example between China and the United States (US), be halted before it gathers real momentum. Again, it may well be too late for this, and European influence is obviously limited. However, research into, and development of, offensive autonomous weapons should not be funded, and indeed should be outright banned, on EU territory. Autonomous weapons select and engage targets without human intervention, and they will act and react on ever shorter timescales, which in turn will make it seem reasonable to transfer more and more human autonomy into these systems themselves. They may, therefore, create military contexts in which relinquishing human control almost entirely seems like the rational choice. Autonomous weapon systems lower the threshold for entering a war, and if both warring parties possess intelligent, autonomous weapon systems there is an increased danger of fast escalation based exclusively on machine-made decisions. In this problem domain, the degree of complexity is even higher than in the context of preventing the development and proliferation of 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 on such prohibitions.

Recommendation 5

For purely defensive military applications (if they are at all conceivable), 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.

IV. A MORATORIUM ON SYNTHETIC PHENOMENOLOGY

It is important that all politicians understand the difference between AI 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, in advanced artificial systems. Potential future artificial subjects of experience currently 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.³ One potential risk is that of dramatically increasing the overall amount of suffering in the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.

For this, I refer readers to an open-access publication of mine, titled 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology'.⁴ The risk that has to be minimised in a rational and evidence-based manner is the risk of an 'explosion of negative phenomenology' (ENP; or simply a 'suffering explosion') in advanced AI and other post-biotic systems. I will here define 'negative phenomenology' as any kind of conscious experience a conscious system would avoid or rather not go through if it had a choice.

On ethical grounds, we should not risk an explosion of conscious suffering – at the very least not before we have a much deeper scientific and philosophical understanding of what both consciousness and suffering really are. As we presently have no good theory of consciousness and no good, hardware-independent theory about what 'suffering' really is, the ENP risk is currently incalculable. It is unethical to run incalculable risks of this magnitude. Therefore, until 2050, there should be a global ban on all research that directly aims at, or indirectly and knowingly risks, the emergence of synthetic phenomenology.

Synthetic phenomenology is only one example of a type of risk to which political institutions have turned out to be systematically blind, typically dismissing such risks as 'mere science fiction'. It is equally important that all politicians understand both the possible interactions amongst specific risks and – given the large number of 'unknown unknowns' in this domain – the fact that there is an ethics of risk-taking itself. This point relates to uncomprehended risks we currently label as 'mid-term', 'long-term', or 'epistemically indeterminate'.

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 on such prohibitions.⁵

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 fields such as philosophy, neuroscience, and computer

³ See T Metzinger, 'Two Principles for Robot Ethics' (2013) in E Hilgendorf and JP Günther (eds), *Robotik und Gesetzgebung*; T Metzinger, 'Suffering' (2017) in K Almqvist and A Haag (eds), *The Return of Consciousness*.

 ⁴ See T Metzinger, 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology' (2021) 8
(1) Journal of Artificial Intelligence and Consciousness 1, 43–66. https://www.worldscientific.com/doi/pdf/10.1142/S270507852150003X.

⁵ 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 S Dehaene, H Lau, and S Kouider, 'What Is Consciousness, and Could Machines Have It?' (2017) 6362 Science 486; MSA Graziano, 'The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness' (2017) 4 Frontiers in Robotics and AI; R Kanai, 'We Need Conscious Robots. How Introspection and Imagination Make Robots Better' (Nautilus, 27 April 2017) http:// nautil.us/issue/47/consciousness/we-need-conscious-robots.

science). Specific topics of relevance 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 most complex, and likely contains the largest number of unexpected problems and 'unknown unknowns'.

V. DANGERS TO SOCIAL COHESION

Advanced AI technology will clearly provide many possibilities for optimising 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 plausible to assume that there are many new, at present unknown, risks with the potential to undermine efforts to sustain social cohesion. It is also reasonable to assume the existence of a larger number of 'unknown unknowns', of AI-related risks that we will discover only by accident and late in the day. 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 automation.⁶ The implementation of AI technology by financially potent stakeholders may lead to a steeper income gradient, increased inequality, and dangerous patterns of social stratification.⁷ 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 via privately owned and autonomously controlled social media aimed at harvesting human attention and 'packaging' it for further use by their customers, or in 'engineering' the formation of political will via Big Nudging strategies and AI-controlled choice architectures that are not transparent to the individual citizens whose behaviour is thus controlled.⁸ Future AI technology will be extremely good at modelling and predictively controlling human behavior - for example by positive reinforcement and indirect suggestions, making compliance with certain norms or the emergence of 'motives' and decision outcomes appear entirely spontaneous and 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, most of these risks to social cohesion are still very likely unknown at present, and we may discover them only by accident. Policymakers must also understand that any technology that can purposefully optimise the intelligibility of its own action for 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.

⁶ See European Parliamentary Research Service "The Ethics of Artificial Intelligence: Issues and Initiatives" (European Parliamentary Research Service, 2020) 6–11.

⁷ A Smith and J Anderson, 'AI, Robotics, and the Future of Jobs' (*Pew Research Center*, 2014) www.pewresearch.org/ internet/wp-content/uploads/sites/9/2014/08/Future-of-AI-Robotics-and-Jobs.pdf.

⁸ For a first set of references, see www.humanetech.com/brain-science.

AI technology is currently a private good. It is the duty 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 are not in existence it will unfold invisibly, for example in decentralised situations. The EU should proactively develop such preventative 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 in 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 to investigate the feasibility of accurately timed initiatives for retraining threatened population strata towards creative 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.

VI. RESEARCH ETHICS

One of the most difficult theoretical problems in this area is the problem of 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 plausibly engendering 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, and 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.⁹ In principle, this could happen in the AI research community as well. Therefore, the EU should certainly 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. Rational, evidence-based identification and

⁹ See FS Collins and AS Fauci, 'NIH Statement on H5N1' (*The NIH Director*, 2012) www.nih.gov/about-nih/who-weare/nih-director/statements/nih-statement-h5n1; and RAM Fouchier and others, 'Pause on Avian Flu Transmission Studies' (2012) Nature 443.

minimisation of risks (including those pertaining to a distant future) ought to be a part of research itself, and scientists should cultivate a proactive attitude to risk, especially if they are likely to be the first to become aware of novel types of risk through their own work. Communication with the public, if needed, should be self-initiated, in the spirit of taking control and acting in advance of a possible future situation, rather than just reacting to criticism by non-experts with some set of pre-existing, formal rules. As *Michael Madary* and I note in our ethical code of conduct for virtual reality, which includes recommendations for good scientific practice: '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.¹⁰

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 in problems of AI technology, autonomous systems, and related fields needs 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 level of academic excellence and professionality.

VII. 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 to optimise the risk/benefit ratio in a timely fashion. The growth of AI exemplifies how powerfully time pressure can constrain rational and evidence-based identification, assessment, and management of emerging risks; creation of ethical guidelines; and implementation of an enforceable set of legal rules. There is a 'pacing problem': Existing governance structures are simply unable to respond to the challenge fast enough; political oversight has already fallen far behind technological evolution.¹¹

I am drawing attention to the current situation not 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

¹⁰ M Madary and T Metzinger, 'Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology' (2016) 3(3) *Frontiers in Robotics and AI* 1, 12.

¹¹ GE Marchant, 'The Growing Gap between Emerging Technologies and the Law' in GE Marchant, BR Allenby, and JR Herkert (eds), *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight* (2011), 19, 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'.

themselves. 'Meta-governance' means just this: A governance of governance equal to 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.¹² *Gary Marchant* and *Wendell Wallach* have, therefore, proposed so-called Governance Coordination Committees (GCCs), a new type of institution providing a mechanism for coordinating and synchronising what they aptly describe as an 'explosion of governance strategies, actions, proposals, and institutions'¹³ 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 analysis and monitoring instrument, and 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.'¹⁴

Many other strategies and governance approaches are, of course, conceivable. However, this is not the place to discuss details. Here, the general point is simply that we can meet the challenge posed by rapid developments in AI and autonomous systems only if we put the question of meta-governance on top of our agenda right from the start. In Europe, the main obstacle to reaching this goal is, of course, 'soft corruption' through the Big Tech industrial lobby in Brussels: There are strong financial incentives and major actors involved in keeping the pacing gap as wide open as possible for as long as possible.¹⁵

Recommendation 15

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

VIII. CONCLUSION

I have proposed that the European Union immediately begin working towards the development of a Global AI Charter, in a multilayered process starting with an AI Charter for the EU 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 15 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

¹² See W Wallach, A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control (2015), 250.

¹³ 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 GE Marchant and W Wallach, 'Coordinating Technology Governance' (2015) 31 Issues in Science and Technology (hereafter Marchant and Wallach, 'Technology Governance').

¹⁴ Marchant and Wallach, 'Technology Governance' (n 14), 47.

¹⁵ For one recent report, see M Bank and others, 'Die Lobbymacht von Big Tech: Wie Google & Co die EU beeinflussen' (Corporate Europe Observatory und LobbyControl e.V., 2021) www.lobbycontrol.de/wp-content/ uploads/Studie_de_Lobbymacht-Big-Tech_31.8.21.pdf.

the problem is that we cannot rely on intuitions, because we must satisfy counterintuitive rationality constraints. Therefore, we also need humility, intellectual honesty, and genuine open-mindedness.

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.¹⁶

¹⁶ Cf. Mannino and others, 'Opportunities and Risks' (n 1).