Ethical AI for the Governance of the Society: Challenges and Opportunities

Nieminen Mika [1[0000-0001-8528-6869]], Gotcheva Nadezhda [1[0000-0002-1449-6647]], Leikas Jaana [1 [0000-0001-7863-7018]] and Koivisto Raija [1]

[1]VTT Technical Research Centre of Finland Ltd., Visiokatu 4, Tampere

mika.nieminen@vtt.fi, nadezhda.gotcheva@vtt.fi, jaana.leikas@vtt.fi, raija.koivisto@vtt.fi

Abstract Artificial Intelligence (AI) technologies are expected to have numerous and diverse social implications that cut deep into our society. Due to AI's specific nature as emergent and constantly evolving generic technology, we need new approaches, methodologies, and processes to govern and steer the utilization of AI technologies both in the public and private sectors. This is both a multi-level and multi-dimensional governance challenge. First, there has to be a shared and coordinated understanding across various social and administrational sectors on how AI is implemented and regulated. Second, good coordination between different levels of governance is crucial. Third, there is a challenge to find a balance between soft and hard governance mechanisms in varying implementation and organizational contexts. This paper presents an overview of a new Strategic Research Council funded project project entitled "Ethical AI for the Governance of the Society" (ETAIROS). The project focuses on studying and co-developing together with stakeholders practical governance approaches, as well as design and technology solutions that help public, private and civil society actors enhance the ethical sustainability of operations in the use of AI. To achieve its ambitious goals, this interdisciplinary endeavour integrates expertise in foresight, ethics, design, machine learning and governance.

Keywords: ethics, artificial intelligence (AI), foresight, design, governance, societal impacts, responsibility

1 Introduction

Artificial Intelligence (AI) technologies are expected to have numerous and diverse implications that cut deep into our society. Various definitions of AI exist and what they have in common is that they usually refer to the increasing capability of machines to perform tasks, which have been conducted traditionally by people. As humans, we

have certain limitations, and AI developers have often focused there, with the presumption that AI is capable to overcome some of these physical, cognitive or other limitations, and to expand the human potential to new horizons. Still, AI technologies and application areas are emerging and there is uncertainty related to many aspects of their design and implementation.

Due to AI's specific nature as constantly evolving generic technologies, we need careful calibration of the existing approaches, methodologies, processes and guiding principles, or development of completely novel ways to govern and steer the utilization of AI technologies both in the public and private sectors. Essentially, this is a multi-level governance challenge. First, there has to be a shared and coordinated understanding across various social and administrational sectors on how AI is implemented and regulated, and, second, coordination between different levels of governance is also necessary. Third, there is a challenge to find an optimal balance between soft and hard governance mechanisms in varying implementation and organizational contexts.

This paper presents an overview of a new project entitled "Ethical AI for the Governance of the Society" (ETAIROS). The project integrates expertise in foresight, ethics, design, machine learning and governance to study and co-develop together with stakeholders practical governance processes and frameworks as well as design and technology solutions that help public, private and civil society actors enhance the ethical sustainability of their operations in the use of AI. The project is implemented by the joint effort of six organizations: Tampere University, VTT Technical Research Centre of Finland Ltd., University of Helsinki, University of Jyväskylä, University of Turku and 4Front. In ETAIROS, we study and co-develop together with stakeholders practical governance processes and frameworks as well as design and technology solutions that help public, private and civil society actors enhance the ethical sustainability of their operations in the use of AI.

The project activities are guided by four scientific objectives: 1) to develop a theoretically and empirically grounded approach for steering the development and use of AI and its societal impacts; 2) to produce a body of novel knowledge on context specific challenges, opportunities and barriers to ethical and responsible use of AI; 3) to deliver tested design and machine learning processes for ethical AI; and 4) to yield empirically justified governance approaches and practices for the use of AI. These contributions are expected to frame the keys to socially sustainable strategic planning, policy and regulation of AI.

2 Background: Current transformation in society and business

AI technologies are revolutionizing our world. AI entities are "digital computers or computer-controlled robots that perform tasks commonly associated with intelligent beings", combining diverse abilities to learn, reason, solve problems, perceive and use language (Copeland 2018). AI technologies vary by the scope and depth of "thinking" they undertake. We already use narrow AI technologies, algorithmic expert systems,

big data, and deep learning in health care, banks, insurance companies, public policy and governance, factory production, security and law enforcement, as well as autonomous vehicles, not to mention social media apps (Grace et al. 2018; O'Neil 2016; OECD 2018). Artificial General Intelligence (AGI) systems that "possess a degree of self-understanding and autonomous self-control" (Goertzel & Pennachin 2007, Saariluoma 2015) are still beyond the horizon.

Even if some of the buzz around AI were hyped, AI has a staggering economic potential (Brynjolfsson & Mcafee 2014, 2017). Forecasts indicate that AI revenues will surge in the coming years. For instance, Tractica research has estimated that the income from AI applications will shift from $643.7 million in 2016 to $36.8 billion by 2025 (Faggella 2018). A recent global survey of business executives reported that some 72% of respondents expected AI to have a significant impact on businesses in the next five years (Ransbotham et al. 2017). Globally, there is a competition who will become the world leader in AI. China, for example, set goals to become one by 2030 (Forbes, 2019). In Europe, spending for AI-based technologies in 2019 has increased 49% over the 2018 figure to reach USD5.2 billion (IDC, 2019). For example, large corporations such as Microsoft just announced their "AI for Good" programme, which aims at "providing technology, resources and expertise to empower those working to solve humanitarian issues and create a more sustainable and accessible world." (Microsoft webpage). This initiative is planned to be run from the UK and it aims at integrating technology, expertise in artificial intelligence and data science with expertise in environmental science, disability needs and humanitarian assistance.

Governments have clearly recognized the AI potential yet some are more vocal as forerunners than others. Nearly all developed nations have AI strategies and compete with each other to support AI development and deployment (Dutton 2018). In 2018 EU set up an AI expert group to prepare a union-wide AI strategy. Simultaneous to its economic promise, AI is likely to have a transformative social and cultural impact. The technologies appear capable of disrupting existing social power structures, industries, even our life as a species (Bostrom, 2014). As a result, recent surveys and studies have tried to gauge the effects AI could have on a wide variety of contexts ranging from politics (Helbing et al, 2017), war (Cummings, 2017) to wealth distribution and employment (e.g. Korinek & Stiglitz 2017, Avent 2016).

United Nations (UN) join forces to ensure AI for Good: in 2019 the UN published a report "UN Activities on Artificial Intelligence", which outlines how AI is being used to fight hunger, ensure food security, mitigate climate change, advance health, and facilitate the transition to smart sustainable cities. It also offers insights into the challenges associated with AI, addressing ethical and human right implications, and so invites all stakeholders, including government, industry, academia and civil society, to consider how best to work together to ensure AI serves as a positive force for humanity and the environment.

3 Challenges and opportunities

Nowadays, AI is surrounded by intense hype. Leikas (2019) reminded that we need to look beyond the hype because real-life examples and in-depth discussions on ethical issues and potential impacts are still insufficient. It is unclear what we are even talking about when we refer to ethics of AI. The problem is, as noted by Leikas (2019) that we easily fall into looking for ethical dilemmas related to AI, while we should be asking how these emerging technologies should be designed and used for good and improving the quality of life. So, when it is asked "whom an autonomous car should be allowed to run over", it is simply a wrong question. The important questions to ask are those which focus on ensuring peace, safety and security of citizens, trust in society, equal availability of services, possibility to be heard, and justifying technology decisions from the perspectives of human dignity, welfare and sustainability.

Society can - and should - adopt AI in a way that maintains or improves the quality of life of citizens. Currently, many expectations are pinned upon AI in different fields of everyday life, yet at the same time, there are a number of ethical questions associated with it. Many of them concern the design of interactions between human and non-human actors that foster trust. These include e.g., the human-machine co-working, the ownership of the used data, and distortions in the data, as well as privacy-preserving and resilient AI. For example, collecting a wealth of personal data for health maintenance when at the same time facing an increase in radical openness give both citizens and decision makers causes of concern. Alike, the promises of AI in work life in terms of autonomous systems as work mates, as well as illustrations of future smart cities with autonomous maintenance and transportation exercise many citizens' minds.

The design and use of AI are inevitably socially and culturally embedded (Kitchin 2017, 18). Research has shown that machine-learning methodologies often give rise to social biases, which derive from the programming choices and data used to train the systems. Such biases may relate e.g. to individuals' gender and ethnic background, and affect their equality of opportunity (Weber 2018). Due to system complexity, it is extremely difficult to identify such biases and develop "debiasing" algorithms (Brynjolfsson & Mcafee 2017).

Challenges have been identified in AI use as well. Algorithmic technologies have been utilized to destabilize democratic processes (Cadwallar & Graham-Harrison 2018) and for purposes of control, which may give rise to societies where panoptic surveillance (Müller 2016) and pervasive scoring (Citron & Pasquale 2014) affect all aspects of our lives. There are even worse dystopias of the end of humanity caused by "singularity" or the devastating consequences criminal use of AI may have (Naudé & Nicola 2018). We seem to be on the cusp of a "new era of widespread algorithmic governance, wherein algorithms will play an ever-increasing role in the exercise of power" (Kitchin 2017, 15). This challenge may be boosted by the fact that machine thinking is not and will unlikely be an imitation or extension of human reasoning (Lake 2017).

Many organizations and public actors, such as governments, have been developing and publishing guidelines for ethical development and use of AI. These initiatives have been triggered by the need to address the potential harms - deliberate or unintentional - AI systems can cause at every stage of their lifecycle harms to individuals, society or

the environment. Perhaps we need to rethink our concept of "lifecycle" and even what it means to be a human in the future. The main challenges associated with the AI systems have generally been described as related to misuse, design that is not thoroughly considered, and unintended negative consequences.

AI holds a lot of opportunities to accelerate innovation and international industry and governmental cooperation in order to tackle key societal challenges of our time. From that perspective, AI can be seen as an important emerging tool to catalyze positive social impact. Decisions taken today and technological solutions designed nowadays may affect current societies and the environment, and also significantly influence the future generations. To benefit from potential opportunities, we need futures thinking and brave action to place humans in the center. Ethics is needed to provide vocabulary and approach for the AI systems developers, implementers, and stakeholders to equip them with values, principles, and practical techniques to mitigate potential current and long term harms associated with AI applications.

4 Moving towards new horizons in developing ethical AI and societal governance

The expected societal impacts of AI concern the public, private and third-sector actors' ethical self-regulation and steering of the society. ETAIROS will advance knowledge on relevant use contexts and specific challenges and opportunities of AI, develop ethical design and assessment frameworks and tools (Leikas et al., 2019), and elaborate general governance principles and practices. For public authorities and private sector, the project produces suggestions and practices for the use, design and governance of AI from the perspective of sustainable, transparent, and inclusive societal development. From the perspective of citizens and civil society, the project increases transparency of the use of AI, general understanding of ethically acceptable AI systems and possibilities for informed public debate and influence.

AI is affecting everyone, and AI applications and autonomous systems are facing huge business expectations and hopes as means to make citizens and societies flourish. To succeed in this, common action and discussion is needed, not only between research and industry but also between citizens, decision makers and companies to domesticate AI in a trustworthy manner in the everyday life of citizens and organisations. To ensure societal impact, ETAIROS will collaborate productively across all sectors by actively engaging all relevant key actors (public authorities, experts, citizens, private sector) to a transparent and well-informed co-innovation process of new practical governance frameworks and tools including regulation suggestions. Concrete use cases are examined to support the formation of shared understanding of the challenges and solutions.

Research in ETAIROS will be executed in two phases: during Phase I (2019-2022) we will study ethical AI development and use, its governance challenges, and develop and pilot frameworks and practical instruments for ethical AI design, use and governance in collaboration with the stakeholders. During Phase II (2022-2025), the frame-

works and practical tools will be refined and finalized on the basis of further experiments, and scaled up to a wider use by public authorities, companies and the third sector.

Interaction activities in ETAIROS aim at co-creating design models and stimulating concrete action for ethical adoption and utilization of AI. ETAIROS brings together researchers, public agencies, policy makers, industry, business community, and civil society actors in a co-creative research and innovation process. The core tools to achieve this goal are the Co-Innovation Forum (CIF) and the Open Dialogue Forum (ODF). The CIF is a forum where practical use case areas of AI are elaborated and co-created together with co-innovation partners. The ODF is open for all relevant actors, including civil society, and supports the ideas of open innovation and open science.

In summary, ETAIROS project is expected to provide novel insights by combining AI design challenges and societal concerns into a single empirical study; anticipating systematically societal impacts of AI development by using established participatory foresight methods; incorporating governance aspects to the inquiry to provide policy and business relevant suggestions and practical solutions; co-innovating societally acceptable and desirable solutions by integrating stakeholders and citizens, and developing tools for screening and enhancing ethical aspects in applications utilizing AI.

Acknowledgements

Project “Ethical AI for the Governance of the Society” (ETAIROS) is funded by Strategic Research Council at the Academy of Finland.

References

1. Avent, R. (2016). The wealth of humans: Work, power, and status in the twenty-first century. NY: St. Martin’s Press.
2. Bostrom, N. (2014) Superintelligence: Paths, dangers, strategies. Oxford University Press.
3. Brynjolfsson, E. & McAfee, A. (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. New York: W.W. Norton & Company.
4. Brynjolfsson, E. & Mcafee, A. (2017) The business of artificial intelligence. What it can-and cannot-do for your organization. Harvard Business Review, 7.
5. Cadwallar, C. & Graham-Harrison E. (2018) How Cambridge Analytica turned Facebook ‘likes’ into a lucrative political tool. The Guardian, March 17.
6. Citron, D. K. & Pasquale, F. A. (2014) The scored society: Due process for automated predictions. Washington Law Review, 89, 2014; U of Maryland Legal Studies Research Paper No. 2014-8.
7. Copeland, B.J. (2018) Artificial intelligence. Encyclopaedia Britannica.
8. Cummings, M. L. (2017) Artificial intelligence and the future of warfare. Research Paper. London: Chatham House.
9. Dutton, T. (2018) An overview of national AI strategies. https://medium.com/
10. Forbes (2019) Artificial intelligence, China and the U.S. – How the U.S. is losing the technology war. https://www.forbes.com/sites/steveandriole/2018/11/09/artificial-intelligence-china-and-the-us-how-the-us-is-losing-the-technologywar/#2dcafacd6195.

11. Faggella, D. (2018) Valuing the artificial intelligence market, graphs and predictions. https://www.techemergence.com/
12. Goertzel B. & Pennachin, C. (Eds.) (2007) Artificial general intelligence. Springer-Verlag: Berlin Heidelberg.
13. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research (AI and Society Track).
14. Helbing, D. et al. (2017) Will democracy survive Big Data and Artificial Intelligence? Scientific American, February 25.
15. IDC (2019) Automation and Customer Experience Needs Will Drive AI Investment to $5 Billion by 2019 Across European Industries. https://www.idc.com/getdoc.jsp?containerId=prEMEA44978619
16. Kitchin R. (2017) Thinking critically about and researching algorithms. Information, Communication & Society, 20(1), 14-29.
17. Korinek, A. & Stiglitz J. (2017) Artificial Intelligence and its implications for income distribution and unemployment. NBER Working Paper.
18. Lake, B., Ullman, T., Tenenbaum, J. & Gershman, S. (2017) Building machines that learn and think like people. Behavioral and Brain Sciences, 40.
19. Leikas, J., Koivisto, R., & Gotcheva, N. (2019) Ethical framework for designing autonomous systems. Journal of Open Innovation: Technology, Market, and Complexity 2019, 5, 18; doi:10.3390/joitmc5010018
20. Leikas, J. (2019) The ethics of AI – what are we even talking about? https://vtt-blog.com/2019/01/16/the-ethics-of-ai-what-are-we-even-talking-about/
21. Microsoft (2019) https://www.microsoft.com/en-gb/ai/ai-for-good [Accessed 1.10.2019]
22. Müller, V. (Ed.). (2016) Risks of Artificial Intelligence. CRC Press, Taylor & Francis Group.
23. Naudé W. & Nicola, D. (2018) The race for an artificial general intelligence: Implications for public policy. UNU-MERIT Working Papers, Maastricht Economic and Social Research institute on Innovation and Technology.
24. O'Neil C. (2016) Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group & Penguin.
25. OECD (2018) AI: Intelligent machines, smart policies: Conference summary. OECD Digital Economy Papers, No. 270, OECD Publishing, Paris.
26. Ransbotham, S., Kiron, D., Gerbert, P. & Reeves, M. (2017) Reshaping business with Artificial Intelligence. Closing the gap between ambition and action. MIT Sloan Management Review.
27. Saariluoma, P. (2015) Four challenges in designing autonomous systems. In: Williams, A., Scharre, P., Mayer, C., Arnold, R., Crootof, R., Anderson, J. & Husniux, A. Autonomous Systems: Issues for Defence Policymakers. NATO.
28. Weber, J. (2018) Auto-management as governance? Predictive analytics in counter-insurgency and marketing. Conference presentation at EASST 2018, July, Lancaster University.