Kerikmäe, T.
PhD, Professor of European Legal Policy and Law & Technology
Department of Law
Tallinn University of Technology
Estonia
Antonov, A.
MSc, Doctoral Candidate
Department of Law
Tallinn University of Technology
Estonia
Shumilo, O.
PhD
Department of Law
Tallinn University of Technology
Estonia

CONCEPTUALIZING ETHICAL CHALLENGES IN DEVELOPMENT AND USAGE OF ARTIFICIAL INTELLIGENCE SYSTEMS: TOWARDS 'ETHICS BY DESIGN'

'*We should work on the competence, capabilities, mechanisms and the supporting institutions that allow us to investigate systematically in moral terms what is designed, developed and produced and identify which values are supported or realised by designs that shape the lives of people. This is what we may call the ideal of 'design for values', 'value-sensitive design' or 'ethics by design*'.

([1]: European Groups on Ethics in Science and New Technologies, 2021)

Discussions on how to leverage the potential of Artificial Intelligence Systems (AI) (*see, definition in*: [2]) for societal good in business and public administration contexts have taken centre stage both in scholarly and policy discourses ([3];[4];[5];[6]), not least amid the COVID-19 pandemic. Digitalization and the ensuing transformative processes, in particular related to development of digital ecosystems [7], challenges variable stakeholders, ranging from CEOs to local policymakers, to rethink traditional business models and delivery of public services [8]. While stakeholders in the private sector are primarily

incentivized to design, train and deploy AI systems for the purpose of empowering consumers, adoption of AI systems at public sector levels takes place in the context of citizen empowerment for public good. This paper argues that reflections on AI policy need be grounded in ethics, fundamental rights *and* value-sensitive design rationales (for an in-depth understanding of value-sensitive design approaches, see: [9]) at *both* private and public sector levels and be open to all democratically minded stakeholders therein, considering the implications wide-scale adoption of AI systems, in particular those classified as high-risk [2], bear for society at large.

To justify this claim, the paper reviews existing AI governance literature through the prism of ethics, fundamental rights *and* value-sensitive design with particular focus on AI development and usage for public services and additionally suggest that key societal benefits of AI systems can only be harnessed if core underlying ethical challenges in development and usage of AI are *incrementally* and *iteratively* put under public scrutiny [10] taking into account that the features of *incremental* and *iterative* policy design, which are characteristic to value-sensitive design approaches, have yet been only weakly adopted at public and business administration levels [9].

Additionally, the theoretically informed argument is premised on the assumption that AI as a general purpose and dual-use technology, metaphorically conceived as a double-edge sword, or in the Estonian context as the mythical figure of *Kratt*, one that is 'devoted to serving its master', but, if not ethically governed, 'can become bad' ([11];[12]), requires a certain method of steering or governance that takes account of context-specificity, thus being *facilitative by design*.

As such, resting on two pillars – *incremental* and *iterative* identification of ethical pitfalls, and *facilitative by design* AI governance – a thorough literature review is conducted to map ethical challenges in development and usage of AI systems. These challenges are thereafter synthesized into five themes in the paper, the output of which culminates into and is illustrated by a conceptual framework, outlining avenues for future research on adoption of value-sensitive design methods into design of AI policies at public and business administration levels.

Essentially, the aim of this research is to contribute to the challenge of *'opening the black box of AI'* to society ([13];[14]), being thus geared towards informing policy discussions on AI at *both* business and public administration levels with ethics, fundamental rights *and* value sensitive design rationales.

List of references

1. European Group on Ethics in Science and New Technologies (2021). Values for the Future: The Role of Ethics in European and Global Governance, Publications Office of the European Union, Luxembourg;

2. EU Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act. EC, COM (2021) 206 (21 April 2021) [Accessed on: 27.10.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206;

3. Dafoe, A. (2018). AI Governance: A Research Agenda. Governance of AI Program, Future of Humanity Institute (27 August 2018) [Accessed on: 27.10.2022] https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf;

4. Floridi, L. et al. (2018). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines*, vol. 28, n. 4: pp. 689–707, DOI: https://doi.org/10.1007/s11023-018-9482-5;

5. Kerikmäe, T., and Pärn-Lee, E. (2021). "Legal Dilemmas of Estonian Artificial Intelligence Strategy: In Between of E-Society and Global Race." *AI & Society*, vol. 36, no. 2: pp. 561-572, DOI: https://doi.org/10.1007/s00146-020-01009-8;

6. Metcalf, K. N., and Kerikmäe, T. (2021). "Machines Are Taking over-Are We Ready?." *The Singapore Academy of Law Journal*, vol. 33: pp. 24-49;

7. Kerikmäe, T. and Ramiro Troitiño, D. (2022). "Introducción. Digitalización de la Unión Europea: repercusiones y expectativas." *Revista CIDOB d'Afers Internacionals*, n. 131: pp. 7-15. DOI: doi.org/10.24241/rcai.2022.131.2.7;

8. Kerikmäe, T.; Hoffmann, T. and Chochia, A. (2018). "Legal Technology for Law Firms: Determining Roadmaps for Innovation." *Croatian International Relations Review*, vol. 24, no. 81: pp. 91-112. DOI: https://doi.org/10.2478/cirr-2018-0005;

9. Friedman, B., and Hendry, D. G. (2019). Value Sensitive Design: Shaping Technology With Moral Imagination. MIT Press, Cambridge;

10. Antonov, A. (2022). "Gestionar la Complejidad: La Contribución de la UE a la Gobernanza de la Inteligencia Artificial." *Revista CIDOB d'Afers Internacionals*, n. 131: pp. 41-68. DOI: https://doi.org/10.24241/rcai.2022.131.41;

11. Kerikmäe, T., Metcalf, K., Hoffmann, T., Minn, M., Liiv, I., Taveter, K., Shumilo, O., Solarte Vasquez, M. C., Antonov, A. (2019). 1st Report on Legal Framework and Analysis Related to Autonomous Intelligent Technologies. (1−11). Riigikantselei;

12. Antonov, A. and Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In: *The EU in the 21st Century*: pp. 135-154. Springer, Cham;

13. Buiten, M. C. (2019). "Towards Intelligent Regulation of Artificial Intelligence." *European Journal of Risk Regulation*, vol. 10, no. 1: pp. 41-59. DOI: https://doi.org/10.1017/err.2019.8;

14. Theodorou, A., and Dignum, V. (2020). "Towards ethical and socio-legal governance in AI." *Nature Machine Intelligence*, vol. 2, no. 1: pp. 10-12. https://doi.org/10.1038/s42256-019-0136-y.