Academy of Finland, Strategic Research Council

Strategic research project: Ethical AI for the Governance of the Society (ETAIROS)

EXECUTIVE SUMMARY

Central idea of the research
Artificial Intelligence (AI) technologies are expected to have numerous and diverse social implications that cut deep into our society. Due to AI technology'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 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. *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.*

To achieve its objectives, ETAIROS 1) combines AI design challenges and societal concerns into a single empirical study; 2) anticipates systematically societal impacts of AI development by using established participatory foresight methods; 3) incorporates governance aspects to the inquiry to provide policy and business relevant suggestions and practical solutions; 4) co-innovates societally acceptable and desirable solutions by integrating stakeholders and citizens, and 5) develops tools for screening and enhancing ethical aspects in applications utilizing AI.

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 yield a body of novel knowledge on context specific challenges, opportunities and barriers to ethical and responsible use of AI; 3) to yield tested design and machine learning processes for ethical AI, 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.

Expected societal impacts
Public, private and third-sector actors' ethical self-regulation and steering of the society will be supported by ETAIROS: we will develop in-depth knowledge on relevant use contexts and specific challenges and opportunities of AI, develop ethical design and assessment frameworks and tools, 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. To ensure societal impact, ETAIROS will actively engage 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) in concrete use cases and support the formation of shared understanding of the challenges and solutions.

Research implementation
The project is organized in seven work packages. WP1 (Foresight) produces scenarios of socio-technical futures and related thematic use cases, insights on the most rapidly developing applications and

contexts, and explores potential societal and public governance impact.WP2 (Ethics) studies theoretically and conceptually relevant ethical aspects of the cases and provides framework for risk assessment.WP3 (Design) studies ethical AI design principles and develops tools for specification-level action design and simulation of AI systems. The tools can interact with ethical machine-learning algorithms developed in WP4 (Machine learning), thus providing a testbed for AI application specifications, applicable in the context of public administration and governance of AI in general. WP5 (Governance) focuses on developing governance methods and regulation models in collaboration with stakeholders.WP6 (Interaction) andWP7 (Management) provide support for constant information exchange and cross-participation across the project.

Interaction

The overall goal of the interaction is to create generic action and design models for ethical adoption and utilization of AI for the governance of the society. 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-innovated 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. The ODF systematically seeks commonalities of the thematic use cases with other areas of social utilization of AI, and seeks to extrapolate the results. In addition, various traditional and social media means are used for effective dissemination and adoption of the findings and results in the society.

Schedule

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