Advances in Reliability, Safety and Security

ESREL 2024
Collection of Extended Abstracts

Building Integrated Legal And Ethical Frameworks For Collaborative Intelligence: Towards More Interdisciplinary Approaches To Human-Robot Collaboration

Naira López Cañellas[a], Aphra Kerr[b], Brian Vaughan[a]

[a]*School of Media, Technological University of Dublin, Ireland*
[b]*Department of Sociology, National University of Ireland Maynooth, Ireland*

Keywords: framework, robotics, artificial intelligence, interdisciplinary research, ethics, governance, regulation

The creation of ethical (non-binding) and legal (binding) instruments to ensure the safety and resilience of AI systems and robotics is an ever-growing practice in sectors that are technology-intensive, given the high pace of innovation and the mistrust that this creates in general audiences and regulatory circles (Cebulla et al., 2022; IEEE, 2019). The motivations behind adopting each of them are different, as are their degree of subjectivity, democratic oversight and conformity requirements (Hagendorff, 2020). Nevertheless, they share the same limitations – there is little guidance on how to operationalise the requirements they outline, and what the role of each stakeholder should be (idem) – which is the gap that this Extended Abstract seeks to close.

This research is based on the document analysis of the most relevant academic and policy legal and ethical publications, from the point of view of companies operating in the domain of AI systems and robotics in Europe. These include:

- from the legal perspective: the General Data Protection Regulation (GDPR, 2016), the Data Act (2023), the Data Governance Act (2022), the Artificial Intelligence Act (2024), and the updated Machinery Directive (2023), as well as their derived international, national and multilateral agreements;
- from the ethical perspective: Recommendation of the Council on Artificial Intelligence of the OECD (2019); Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence of the European Commission (2019); Guidelines for Ethically-Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019); and AI Ethics Guidelines of the Council of Europe's Ad Hoc Committee On Artificial Intelligence (CAHAI, 2020).

Their analyses allow to find where there is the most overarching agreement, and to pool together their unique inputs, in order to create an original framework that is as comprehensive and integrative as possible. The Extended Abstract then continues with a literature review of the state-of-the-art discussions taking place in academic and policy circles regarding the blind spots, limitations, impracticalities and avenues for improvement these pieces of legislation and ethical guidelines have. By taking the experience of the CISC project as a case study, we produce recommendations that provide guidance on how to bridge the current gap between theory and implementation of both ethics and law, particularly tailored to sectors that rely on critical safety systems.

Building An Integrated CI Framework: A Step-By-Step Guide

The research unveils there is a clear set of 7 principles that, despite small alterations and adjustments, seem to be consistently present in almost, if not all, the ethical guidelines mentioned above. These are:

- human agency, autonomy, oversight, liberty and dignity;
- technical robustness and safety;
- privacy and data governance (beyond GDPR);
- transparency, explainability and traceability;
- diversity, non-discrimination and fairness;
- societal and environmental well-being, and

- accountability.

However, the literature review reveals some of these, such as transparency or accountability, are given a much more central role in both discussion and implementation than others, e.g. societal and environmental well-being (Hagendorff, 2020). Similarly, several authors denounce the lack of engagement of disempowered stakeholders in the debate and decision-making regarding many of the regulations and ethical guidelines mentioned above. They highlight the overrepresentation of big companies and little presence of civil society organisations (CSOs), small and medium enterprises (SMEs), and interest groups (IGs) representing the environmental, consumer or worker perspectives (FRA, 2019). And just as frequently, the lack of adaptability of most of these guidelines to a diverse geographical and sectoral contexts, due to their little clarity, accessibility, practicality and/or attention to detail (CAHAI, 2020).

Instead, Hansson & Fenet-Chantereau (2020) suggest that, for any of these governance instruments to fulfil their purpose, there must be greater specification of the objectives of the AI/robotic systems deployed; as well as of the conditions, actors and validation procedures required for them to work as intended throughout their whole lifecycle.

This Extended Abstract aims to build a comprehensive ethical and legal framework adapted to CISC's Live Labs, experimentation environments which test different collaborative intelligence settings through monitoring the muscle and brain activity of an operator.

Some of the risks that have been found in the CISC Live Labs include the potential for invasive surveillance practices (Cebulla et al. 2022), misleading, abusive or negligent use and treatment of data (FRA, 2019), amplification of various forms of discrimination and unequal power dynamics (Cebulla et al., 2022), forceful or unsafe deployment of insufficiently-tested/tailored technologies (Hansson & Fenet-Chantereau, 2020), and unforeseen effects in the quality and security of employment (idem).

Given these risks, these Extended Abstract's recommendations include:

- the need for a greater representation of CSOs, SMEs and IGs in all decision-making processes involved in the production of legal and ethical documents regarding AI systems and robotics, whether happening in the public (EU institutions) or private sectors (e.g. standard-setting bodies);
- going from a *code of conduct* approach towards an *ethics-first* organisations culture that involves e.g. internal ethics review boards and carrying out regular human rights, social and/or environmental impact assessments;
- substituting technical fixes for detail-oriented, context-sensitive and subjectivity-aware ethics;
- prioritising investment in solutions that enhance, amplify and complement human capabilities, relationships and decision-making abilities, instead of those that substitute, challenge or contest them;
- increasing the requirements in the areas of data loss or misuse, cybersecurity vulnerabilities, the unfit/disproportionate collection of operators' data, the justification of algorithmic decisions, etc.

Further research is ongoing in the framework of the CISC project and beyond, in order to continue to monitor present and future clashes between industrial and academic CI settings and the legislation and ethical guidelines designed to address them.

Acknowledgements

This work has been done within the Collaborative Intelligence for Safety-Critical Systems project (CISC). The CISC project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 955901.

References

CAHAI. 2020. AI Ethics Guidelines: European and Global Perspectives. The Council of Europe's Ad Hoc Committee On Artificial Intelligence. https://rm.coe.int/cahai-2020-07-fin-en-report-ienca-vayena/16809eccac.

Cebulla, A. et al. 2022. Applying ethics to AI in the workplace: the design of a scorecard for Australian workplace health and safety. AI & society, 1–17. https://doi.org/10.1007/s00146-022-01460-9.

FRA. 2019. The General Data Protection Regulation – One year on – Civil society: awareness, opportunities and challenges, Publications Office. European Union Agency for Fundamental Rights. https://data.europa.eu/doi/10.2811/538633.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines 30, 99–120.

Hansson, M., Fenet-Chantereau, S. 2020. SIENNA D2.7: Proposal for an ethical framework for the assessment of genomics technologies and for research in genetics and genomics (V1.3). Zenodo. https://doi.org/10.5281/zenodo.7266806.

IEEE. 2019. Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Institute of Electrical and Electronics Engineers (IEEE). First edition. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf .