Imagining the AI Landscape after the AI Act, Third Edition*

Desara Dushi[1,†], Francesca Naretto[2,*,†] and Francesca Pratesi[3,†]

[1]*Vrije Universiteit Brussel, Belgium*

[2]*Dept. of Computer Science, University of Pisa, Italy*

[3]*CNR, Pisa, Italy*

Abstract

We provide a summary of the third Workshop on Imagining the AI Landscape after the AI Act (IAIL 2024), co-located with the 3rd International Conference on Hybrid Human-Artificial Intelligence (HHAI 2024), held on June 10, 2024 in Malmö, Sweden.

Keywords

IAIL, AI Act, Artificial Intelligence, EU, regulation, technology, law, ethics

1. Introduction

After long debates and several amendments to the initial draft, the AI Act was finally adopted and published in the Official Journal on 12 July 2024 becoming the world's first general legislation on artificial intelligence. It aims to provide a framework for the development, placing on the market and use of artificial intelligence (AI) systems, which may pose risks to health, safety and fundamental rights. The AI Act started its gradual entry into force on 1st of August. Its entry into force will take place gradually, encompassing several stages. The first rules entering into force are the ones on prohibited AI systems which will start applying 6 months after entry into force, in February 2025. Secondly, rules for general-purpose AI models will start applying 12 months after entry into force, hence in August 2025. Thirdly, 24 months after entry into force, the rules on high-risk AI systems in annex III become applicable (AI systems in the fields of biometrics, critical infrastructure, education, employment, access to essential private and public services, law enforcement, migration and border control management, democratic processes and the administration of justice). And lastly, 36 months after entry into force, the rules on high-risk AI systems listed in annex I become applicable (toys, radio equipment, in vitro diagnostic medical devices, civil aviation safety, agricultural vehicles, etc.). The entry into application will be based on "harmonised standards" at European level which must define precisely the requirements applicable to the AI systems concerned. The AI Act follows a risk-based approach by classifying AI systems into four levels: unacceptable risk which led to a list of prohibited practices, deemed contrary to the values and fundamental rights of the EU; high risk AI systems, subject to detailed requirements (conformity assessments, technical documentation, risk management mechanisms, fundamental rights impact assessment); specific transparency risk, following a set of transparency obligations; and minimal risk, for all other AI systems, without any specific obligations. Moreover, the AI Act also provides a framework for a new category of so-called general-purpose AI models, in particular in the field of generative AI. These models are defined by their ability to serve a large number of tasks making it difficult to classify them in the previous categories. For this category, the AI Act provides for several levels of obligations, ranging from minimum transparency and documentation measures to an in-depth assessment and the implementation of systemic risk mitigation measures that some of these models might entail, in particular, because of their risks towards major accidents, misuse potential, the spread

HHAI 2024: Hybrid Human AI Systems for the Social Good, June 10-14, 2024, Malmö, Sweden

*Corresponding author.

†These authors contributed equally.

desara.dushi@vub.be (D. Dushi); francesca.naretto@unipi.it (F. Naretto); francesca.pratesi@isti.cnr.it (F. Pratesi)

0000-0001-9714-8033 (D. Dushi); 0000-0003-1301-7787 (F. Naretto); 0000-0002-4260-4473 (F. Pratesi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

of harmful biases and discriminatory effects against certain persons, etc. The AI Act entails a two-level governance structure: European and national level. At European level, the AI board, comprised of representatives from each Member State and the European Data Protection Supervisor (EDPS) as observer, will ensure consistent application of the AI Act. The AI Board will be informed in its choices by an advisory forum and a scientific panel of independent experts. In addition, an AI office in the European Commission will supervise general-purpose AI models. At the national level, the AI Act provides for the designation of one or more competent authorities to assume the role of market surveillance authority. Despite being an EU legislation, the AI Act has an extraterritorial scope, applying to all AI systems that have an impact on European citizens, regardless of the location of the AI system's provider and deployer. Such a broad application will undoubtedly have a significant impact in the EU and beyond. Almost in parallel with the AI Act, on 17 May 2024, after two years of drafting and negotiation, the Council of Europe (CoE) adopted its Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (the CoE Framework Convention), the world's first binding AI treaty. It will be open for signature from 5 September 2024. The EU and CoE have been actively collaborating in the discussions for the draft texts of the two legal documents, ensuring consistency in the terminology and principles within the two texts. The CoE Framework and the EU AI Act complement each other, being an important step towards global AI regulation. The purpose of IAIL 2024 was to explore with young and established experts the effect of the AI Act on technological development in the EU and how it will impact non-EU based developers operating in the EU. It also aimed at analyzing how can the legal requirements of the AI Act be operationalized, to what extent the AI Act protects the fundamental rights of end-users, and much more. Topics of interest included but were not limited to:

- The AI Act and future technologies
- Applications of AI in the legal domain
- Ethical and legal issues of AI technology and its application
- Dataset quality evaluation
- AI and human oversight
- AI and human autonomy
- Accountability and Liability of AI
- Algorithmic bias, discrimination, and inequality
- Fairness by design
- AI and trust
- Transparent AI
- Explainable by design
- Explainability metrics and evaluation
- AI and human rights
- The impact of AI and automatic decision-making on rule of law
- Privacy by design
- AI risk assessment
- AI certification
- Safety, reliance and trust in human-AI interactions
- Human-in-the-loop paradigm
- Federated learning
- Contestability of AI output
- Generative AI

2. Organization

2.1. Workshop Chairs

- Desara Dushi, Vrije Universiteit Brussel (Belgium)

- Francesca Naretto, Scuola Normale Superiore (Italy) and Computer Science Department - University of Pisa (Italy)
- Francesca Pratesi, Institute of Information Science and Technologies - National Research Council (Italy)

2.2. Program Committee

- Costanza Alfieri - University of L'Aquila
- Denise Amram - Scuola Superiore Sant'Anna
- Valeria Caforio - Università Bocconi
- Federica Casarosa - Scuola Superiore Sant'Anna
- Olga Gkotsopoulou - Vrije Universiteit Brussel
- Rami Haffar - Universitat Rovira i Virgili
- Iulia Lefter - Delft University of Technology
- Irina Lishchuk - Leibniz Universität Hannover
- Giorgia Pozzi - Delft University of Technology
- Clara Punzi -Scuola Normale Superiore
- David van Putten - Erasmus University Rotterdam
- Giulia Schneider - Catholic University of the Sacred Heart
- Mattia Setsu - University of Pisa
- Francesco Spinnato - University of Pisa

3. Summary of the workshop

The workshop was highly interdisciplinary and brought together researchers from different backgrounds. The workshop consisted of two keynote speeches, one from Katja de Vries, Associate Professor in public law at Uppsala University, Sweden, and one from Yves-Alexandre de Montjoye, Associate Professor of Applied Mathematics and Computer Science at Imperial College, London, and two sessions of paper presentations with a QA. Regarding the paper presented, we had a contribution from Doh Miriam (University of Mons and The Free University of Bruxelles, Belgium) and Karagianni Anastasia (Vrije Universiteit Brussel, Belgium), titled "My kind of woman: Analysing Gender Stereotypes in AI through The Averageness Theory and EU Law". Following, Nimrod Mike (University of Budapest, Hungary) presented a paper titled "Global Perspectives on AI Governance: A Comparative Overview". Last, but not least, Roberta Savella (ISTI-CNR, Italy) presented a paper titled "The need for a new 'right to refuse' the results of emotion recognition AI". Thanks to the diverse contributions and the two inspiring keynotes, we were able to explore various topics and engage in meaningful discussions, exchanging ideas across multiple research fields, in particular law, computer science, sociology and philosophy. Katja de Vries conducted an analysis on technological advancements in the context of AI and data analytics from the point of view of the legal rights, the privacy implications and the discriminatory processes implicitly conducted in the generation of images and videos from generative AI. Yves-Alexandre de Montjoye discussed the topic of Ethical AI with a focus on privacy, concerning the risks posed by privacy attacks on machine learning models. He explored the balance between privacy and other ethical values required for trustworthy AI systems, as well as the legal implications of the AI Act and its relationship with the General Data Protection Regulation (GDPR). In particular, we discuss the fact that in recent years several regulatory frameworks have been proposed to regulate AI. Not only the AI Act, which was the primary focus of our workshop, but also laws from the United States, China, Japan and many others. Across all of these documents, the principles of transparency, fairness, and accountability are consistently emphasized as essential to achieving trustworthy AI systems. However, even if there is a general acknowledgment about the good work done so far from the different countries, there are still open questions and problems related to some AI applications. For instance, Roberta Savella highlighted that the use of emotion recognition techniques raises concerns that the AI Act does not offer sufficient safeguards for the rights and freedoms of individuals. The issue resides in the fact that, apart from educational institutions and workplaces, facial emotion recognition can be used for lawful purposes as long as it adheres to the obligations imposed on high-risk systems. However, this provision conflicts with Article 22 of the General Data Protection Regulation (GDPR). Therefore, it is important to consider the possibility of granting individuals the right to refuse the use of these technologies when they result in legal consequences. While there are still unresolved questions about the AI Act and its implementation, it remains the only framework that adopts a risk-based approach. This is an important foundation for regulating AI systems without limiting innovation and research in the field. In contrast, the United States relies on a mix of federal and state laws, coupled with industry self-regulation, while China primarily focuses on administrative laws aimed at the deployment of efficient AI systems. These approaches are important and highlight different interests of different countries, but it is necessary to find a shared framework for advancing trustworthy AI. In fact, even if values like transparency, fairness, and accountability are common across most regulations, they alone are not enough to establish a unified global standard. This is why international collaborations, such as the Global Partnership on AI, which aims to harmonize AI governance, are becoming increasingly important. As stated by Mike Nimrod, such initiatives are crucial for fostering the development of a trustworthy AI system. Another key point explored during our workshop, thanks to the presentation of Miriam Doh and Anastasia Karagianni, was the issue of bias in various contexts. The growing use of Large Language Models (LLMs), trained on massive datasets from diverse sources, has raised concerns for several reasons. Primarily, LLMs pose threats to user privacy and complicate accountability. However, one of the most pressing issue from our perspective is bias. In fact, since LLMs have been trained on human data, they are susceptible to inheriting human biases, which can lead to undesirable outcomes. In particular, it has been demonstrated that AI systems, similar to humans, show variations in gender classification accuracy based on the perceived attractiveness of the individual, reflecting human biases. This is just one example of the many problems caused by bias in LLMs, but it highlights the severity of the issue. This challenge must be addressed more comprehensively in future legislation to mitigate the risks posed by biased AI systems.

3.1. Sunbmissions

The Program Committee (PC) received a total of 7 submissions. Each paper was peer-reviewed by at least three PC members, following a double-blind reviewing process. The committee decided to accept 3 papers.

3.2. Detailed Program

The IAIL 2024 program was organized in welcome and final remarks sessions, two invited talks and two paper presentation sessions. The keynotes were in the morning, with an engaged discussion after each one. Following, during the afternoon we had the papers presentation. The papers presentation sessions followed a highly interactive format. They were structured into short presentations with ample room for questions and comments.

4. Summary of the presentations

The workshop had diverse contributions and two keynotes, allowing for discussions across law, computer science, sociology, and philosophy on The AI Act. Several regulatory frameworks, including the AI Act, U.S., China, and Japan's laws, emphasize transparency, fairness, and accountability as crucial to trustworthy AI. However, concerns remain, such as the lack of sufficient safeguards for facial recognition under the AI Act, and conflicts with the GDPR, as highlighted by the contribution of Roberta Savella. While the AI Act adopts a risk-based approach, the U.S. relies on federal and state laws, and China on administrative rules. Global collaborations, like the Global Partnership on AI, are seen as essential to harmonizing AI governance, as reported in the work of Mike Nimrod. During the paper presentation, it was also examined the topic of bias in Large Language Models (LLMs). This topic is only marginally considered in the AI Act, but it is of utmost importance due to its several ethical implications. In particular, there is the problem of potential human biases injected into the LLMs during the training. As an example, it has been shown that there are biases related to gender classification based on attractiveness. This underlines the need for more robust regulatory measures to address bias in AI systems, as highlighted in the work of Miriam Doh and Anastasia Karagianni.

5. Conclusion and Remarks

From the discussion carried out in the IAIL 2024 workshop, it appears evident that multidisciplinarity is a key point for the effectiveness of the EU legal and ethical framework. The workshop itself is a small evidence of the productive results arising from the dialogue of scholars in the different disciplines, having different approaches and motivations. Engaging in conversations and collaborations on human rights is the main goal that needs to be pursued in Europe and hopefully even beyond. Other aspects are the importance of taking particular care of generative AI, the problem of many hands in dealing with the accountability principle, and the need for concrete steps to operationalize the AIA. The papers highlighted the importance and the strength of having a uniform EU legal and ethical framework, as well as the need for a global collaboration to better shape the regulations for achieving trustworthy AI systems.

6. Acknowledgments

This workshop was partially supported by the following:

- SoBigData++ (GA n. 871042) - "European Integrated Infrastructure for Social Mining and Big Data Analytics" (https://plusplus.sobigdata.eu).
- The TAILOR network (GA n. 952215) - Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization (https://tailor-network.eu).
- HumanE-AI-Net (GA n. 952026) - European network of Human-Centered Artificial Intelligence (https://www.humane-ai.eu/).
- ALTEP-DP (SRP54) – Articulating Law, Technology, Ethics and Politics (https://lsts.research.vub.be/ALTEP_DP)
- TANGO (101120763) - European Union's Horizon Europe research and innovation programme (https://tango-horizon.eu/category/news-events/page/2/)

Authors would like to thank the HHAI 2024 workshop chairs and organization for providing an excellent framework for IAIL 2024.