Responsible Artificial Intelligence Systems: From trustworthiness to governance

Framcisco Herrera
Dept. Computer Science and Artificial Intellience, University of Granada
Andalusian Research Institute on Data Science and Computational Intellience
Granada, Spain
herrera@decsai.ugr.es

Abstract— In this plenary talk a present the general framework for designing Responsible Artificial Intelligence Systems, AI systems ensuring auditability and accountability during its design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used. We will discuss two fundamental aspects in this kind of AI systems respecting current regulations, such as trustworthiness and governance, and the path between them.

Keywords—Artificial Intelligence, responsible AI, trustworthiness, AI governance

I. Introduction. Responsible artificial intelligence systems

2023 has been the year of debate on the regulation of artificial intelligence (AI) by governments and institutions such as the United Nations, together with highlight catastrophic statements from business leaders. Among the declarations we find the request for a giant pause or different statements about the high risks of AI. Let us remember the phrase San Altman's statement in a Committee of the US Congress "My worst fear is that this technology will go wrong, and if it goes wrong it can go very wrong." The following two papers provide a thorough analysis of the AI risks [1,2].

In a recent Toju Duke interview [3] (former director of responsible AI program at Google), she focused attention on the current problem "that artificial intelligence amplifies systemic injustices that we should have already eliminated". She highlighted that the debate about this technology has focused on whether humanity will be in danger tomorrow, when the problem is that today it already discriminates against the population. The current dangers can be grouped into three major real scenarios that we must address to regulate and govern AI developing responsible AI systems. They can be summarized as:

- the rise and amplification of misinformation (such as the prevalence of false and malicious content on social platforms),
- biases that reinforce inequalities (gender, ethnicity people discrimination, digital and poverty divide, social credit system/scoring, …)
- the breaking of all the limits of privacy to collect the data that feeds the algorithm and that remains hidden.

In this context, a technical approach to AI emerged, called *trustworthy AI* [4,5]. It is an AI paradigm that encompasses all technical approaches and tools to develop, deploy and use safe, legal and ethical AI systems. It is composed of three pillars and seven requirements (from the European guidelines) [4]. They are the legal, ethical, and robustness pillars; and the technical requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability. On January 2023, NIST released the AI Risk Management Framework [6], including a quite similar list of *trustworthy AI* requirements. We can highlight "secure & resilient", "explainable and interpretable" and "valid & reliable" AI systems, pointing out that approaches which enhance AI trustworthiness can reduce negative AI risks.

It should be noted that he adoption of *trustworthy AI* [7] in the form of practical frameworks is not yet a reality, it is very underdeveloped and conceptual models to materialize this concept are just being born, and are far from common practice (see, for example, the TAII framework [9] and Wasabi conceptual model [10]). The key element in this context must be the concept of "*responsible AI system"* [5]:

Definition*. "A responsible AI system is an AI systems that* requires ensuring auditability and accountability during its design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used."

The implementation of responsible AI can help reduce AI bias, create more transparent AI systems and increase end-user trust in those systems.

II. From trustworthiness to governance

The adoption of trustworthy AI as term has been a milestone in the development of responsible AI. Other terms have been used during these years with similar meanings but the consequent differences in the terms themselves, mong these terms we find "ethical AI", "Ai for good", "beneficial AI", "Responsible AI" or "reliable AI".

The next step in the development process is the AI governance. We trust AI, but we must design the governance at the regulation scenario. Returning to regulation there is growing demand to frame AI regulations to minimize the risks to public safety and preserve human rights while at the same time enabling a flexible and innovative environment [11]. We highlight four initiatives throughout 2023:

- United Nations has created an AI Advisor Body with a clear roadmap, to have a document on governance in mid-2024, for which there is already an interim report "Governing AI for Humanity" (dec. 2023) [12].

979-8-3503-4859-0/DATE24/© 2024 EDAA

- The president of the United States has signed an executive order to regulate the development of artificial Intelligence (October 30, 2023), called "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." [13].

- The AI Safety Summit celebrated in UK, November 1-2, 2023, brought together 30 countries, including China and USA among others, and a Policy paper "The Bletchley Declaration by Countries attending" [14] was published. WE can summarize it with the following short sentences: "*In the context of our cooperation, and to inform action at the national and international levels, our agenda for addressing frontier AI risk will focus on:*
  identifying AI safety risks of shared concern, …
  building respective risk-based policies across our countries to ensure safety in light of such risks, …"
- The European Parliament and Council reach a consensus to approve the AI act [15], regulatory European framework based on risk levels. Two fundamental aspects to highlight, according to the Commission are: "The AI Act transposes European values to a new era. By focusing regulation on identifiable risks, today's agreement will foster responsible innovation in Europe." "By guaranteeing the safety and fundamental rights of people and businesses, it will support the development, deployment and take-up of trustworthy AI in the EU. Our AI Act will make a substantial contribution to the development of global rules and principles for human-centric AI."

In this context, the AI governance emerge as fundamental area, to pilot the regulation and to manage the legal framework for ensuring trustworthy AI technologies are developed with the goal of helping humanity navigate the adoption and use of *responsible AI systems* in safe, ethical and responsible ways. From a holistic point of view, AI governance 360º regulates and manages the safe AI lifecicle, including the continuous monitoring of AI systems. It closes the gap that exists between accountability and ethics in technological AI advancement, piloting a responsible regulation, towards the design of *responsible AI systems*.

We can find two essential elements for AI governance consolidation are along the IA lifecicle:

Auditability. It is becoming increasingly important when standards are being materialized regarding all *trustworthy AI* technical requirements. In terms of particular tools for auditing, especially when the *AI system* interacts with the user, grading schemes adapted to the use case are needed to validate an *intelligent system.*

AI safety. It is an interdisciplinary field concerned with preventing accidents, misuse, or other harmful consequences that could result from AI systems. It encompasses machine ethics and AI alignment, which aim to make AI systems moral and beneficial, and robustness technical problems, adversarial robustness, detecting malicious use, attacks and backdoors, … It includes monitoring systems and alignment. Beyond AI research, it involves developing norms and policies that promote safety [16].

In summary, *responsible AI systems* must be auditable to guarantee their adaptation to the problem context and ethical and legal requirements, and must include in their governance all the elements that guarantee their robustness and security throughout their lifecycle, including the three well-known stages: i) data collection & processing; ii) model training, and iii) model deployment. Following this AI safety approach, *Responsible AI systems* will only reach their full potential when trust can be established in each stage of its lifecycle, from design to development, deployment and use.

III. Conclusions

In short, this approach can allow us develop *responsible AI systems* to reach trustworthiness based on AI governance 360º, covering the bridge from trustworthiness to governance.

References

[1] Critch, A., Russell, S. (2023). TASRA: a Taxonomy and Analysis of Spcietal-Scale Risks from AI. arXiv 2306.06924 (version v2)

[2] Hendrycks, D., Mazeika, M., Woodside, T. (2023) A overview of Catastrophic AI Risks. arXiv:2306.12001 (Version v2)

[3] Toju Duke interview. EL Pais. https://elpais.com/tecnologia/2023-10-20/toju-duke-la-inteligencia-artificial-amplifica-injusticias-sistemicas-que-ya-deberiamos-haber-eliminado.html (in spanish), accessed on January 6th, 2024.

[4] The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. High-Level Expert Group on Artificial Intelligence. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

[5] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., et al. (2023), Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, p.101896.

[6] Artificial Intelligence Risk Management Framework (AI RMF 1.0). Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[7] Fernández-Llorca, D., Gómez, E (2023). Trustworthy Artificial Intelligence Requirements in the Autonomous Driving Domain. *Computer*, vol. 56, no. 2, pp. 29-39.

[8] Li, B., Qi, P., Liu, B., et al. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1-46

[9] Baker-Brunnbauer, J. (2021). TAII Framework for Trustworthy AI Systems. ROBONOMICS: The Journal of the Automated Economy, 2, 17, Available at SSRN: https://ssrn.com/abstract=3914105

[10] Singh, A. M., & Singh, M. P. (2023). Wasabi: A conceptual model for trustworthy artificial intelligence. Computer, 56(2), 20-28.

[11] Almeida, V, Mendes L. S. and Doneda, D. (2023), On the Development of AI Governance Frameworks," in *IEEE Internet Computing*, vol. 27, no. 1, pp. 70-74

[12] AI Advisory Body of United Nations (2023). Interim Report: Governing AI for Humanity. Available at: https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf, accessed on Jan. 6th, 2024.

[13] The White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Available at: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, accessed on Jan 6th, 2024.

[14] [Policy paper. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 Nov. 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, accessed on Jan 6th, 2024.

[15] Artificial Intelligence Act. Available at: https://artificialintelligenceact.eu/, accessed on January 6th, 2024.

[16] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022), Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916 (Version v5).*