Toward Responsible Artificial Intelligence systems: safety and trustworthiness

Francisco Herrera

Dept. Computer Sciences and Artificial Intelligence
Andalusian Research Institute on Data Science and Computational Intelligence
University of Granada
herrera@decsai.ugr.es

Summary. This short paper associated to the extended invited reading introduces two key concepts essential to artificial intelligence (AI), the area of *trustworthy AI* and the concept of *responsible AI systems*, fundamental to understand the technological, ethical and legal context of the current framework of debate and regulation of AI. The aim is to understand their dimension and their interrelation with the rest of the elements involved in the regulation and auditability of AI algorithms in order to achieve safe and trusted AI. We highlight concepts in bold in order to fix the moment when they are described in context.

Extended abstract

Artificial Intelligence (AI) has matured as a technology, AI has quietly entered our lives, and it has taken a giant leap in the last year. Image generative AI models such as Stable Diffusion, Midjourney or Dall-E 2, or the latest evolutions of large language models such as GPT-4 or Bart, have meant that AI has gone, in just a few months, practically from science fiction to being an essential part of the daily lives of hundreds of millions of people around the world.

This emergence goes hand in hand with a growing global debate on the ethical dimension of AI. Concerns arise about its impact on data privacy, fundamental rights and protection against discrimination in automated decisions, or the continued presence of fake videos and images. While some risks of AI, such as the potential for automated decisions harmful to certain vulnerable groups, are relatively well known, there are other less obvious risks, such as hidden biases that may arise from the data used in its training or the vulnerability of AI systems to adversarial attacks.

This whole scenario raises the need to establish responsible, fair, inclusive, trustworthy, safe and transparent frameworks. Before defining precisely these concepts, let's delve into the current state of the AI regulation.

The AI Act draft proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on AI [AIA23] is the first attempt to enact a horizontal AI regulation. The proposed legal framework focuses on the specific use of AI systems. The European Commission proposes to establish a technology-neutral definition of AI systems in EU legislation and defines a classification for AI systems with different requirements and obligations tailored to a "risk-based approach", where the obligations for an AI system are proportionate to the level of risk that it poses.

In this context, a technical approach to AI emerges, called ***trustworthy AI*** [Dia23]. It is a systemic approach that acts as prerequisite for people and societies to develop, deploy and use AI systems. It is composed of three pillars and seven requirements: the legal, ethical, and robustness pillars; and the following 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 top of this, it is necessary to consider a holistic view of trustworthy IA, as outlined in [Dia23], by bridging the gap between theory and practice. This holistic view offered aims to ultimately highlight the importance of all these elements in the development and integration of human-centered AI-based systems into the everyday life of humans, in a natural and sustainable way. We introduce shortly the two fundamental sides, theory and practice:

- Theory: ethical principles, philosophical approach to AI ethics, and key technical requirements (always technically discarding explainability or privacy-based algorithms such as federated learning with multiple private information sources, algorithmic fairness, among others),
- Practice: that revolves around regulation based on risk levels, and the design of intelligent systems that follow this regulation from a legal and ethical point of view. These systems are called "*responsible AI systems"*, and we focus our attention on them in this reading.

It should be noted that he adoption of *trustworthy AI* [Kau22, Li23] 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 [Bak21] and Wasabi conceptual model [Sin23]).

The term ***responsible AI*** has been widely used quite as a synonym of *trustworthy AI*. However, it is necessary to make an explicit statement on the similarities and differences that can be established between trustworthy and responsible AI. The main aspects that make such concepts differ from each other is that *responsible AI* emphasizes the ethical and legal use of an AI-based system, its auditability, accountability, and liability, whereas *trustworthy IA* also consider technological requirements like explainability, robustness, algorithmic fairness, …

To fix the concepts, when referring to *responsibility* over a certain task, the person in charge of the task assumes the consequences of his/her actions/decisions to undertake the task, whether they result to be eventually right or wrong. When translating this concept of responsibility to AI-based systems, decisions issued by the system in question must be accountable, legally compliant, and ethical.

Responsible AI is an area of AI governance, developing AI from both an ethical and legal point of view. The key element in this context is the concept of "***Responsible AI system****":*

> "*It 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." [Dia23].

The implementation of responsible AI can help reduce AI bias, create more transparent AI systems and increase end-user trust in those systems. We introduce shortly the two fundamental features:

- **Auditability** 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.*

- **Accountability** establishes the liability of decisions derived from the *AI system*'s output, once its compliance with the regulations, guidelines and specifications imposed by the application for which it is designed has been audited. Again, accountability may comprise different levels of compliance with the requirements for trustworthy AI defined previously.

It is important to pay attention to auditability (ex-ante) versus accountability (post-hoc) in *intelligent systems* analysis (see Figure 1 for a graphical global vision of both approaches, *auditability & conformity* versus *monitoring & accountability*). The challenge is in the design of auditability methodologies and metrics and accountability monitoring methodologies.

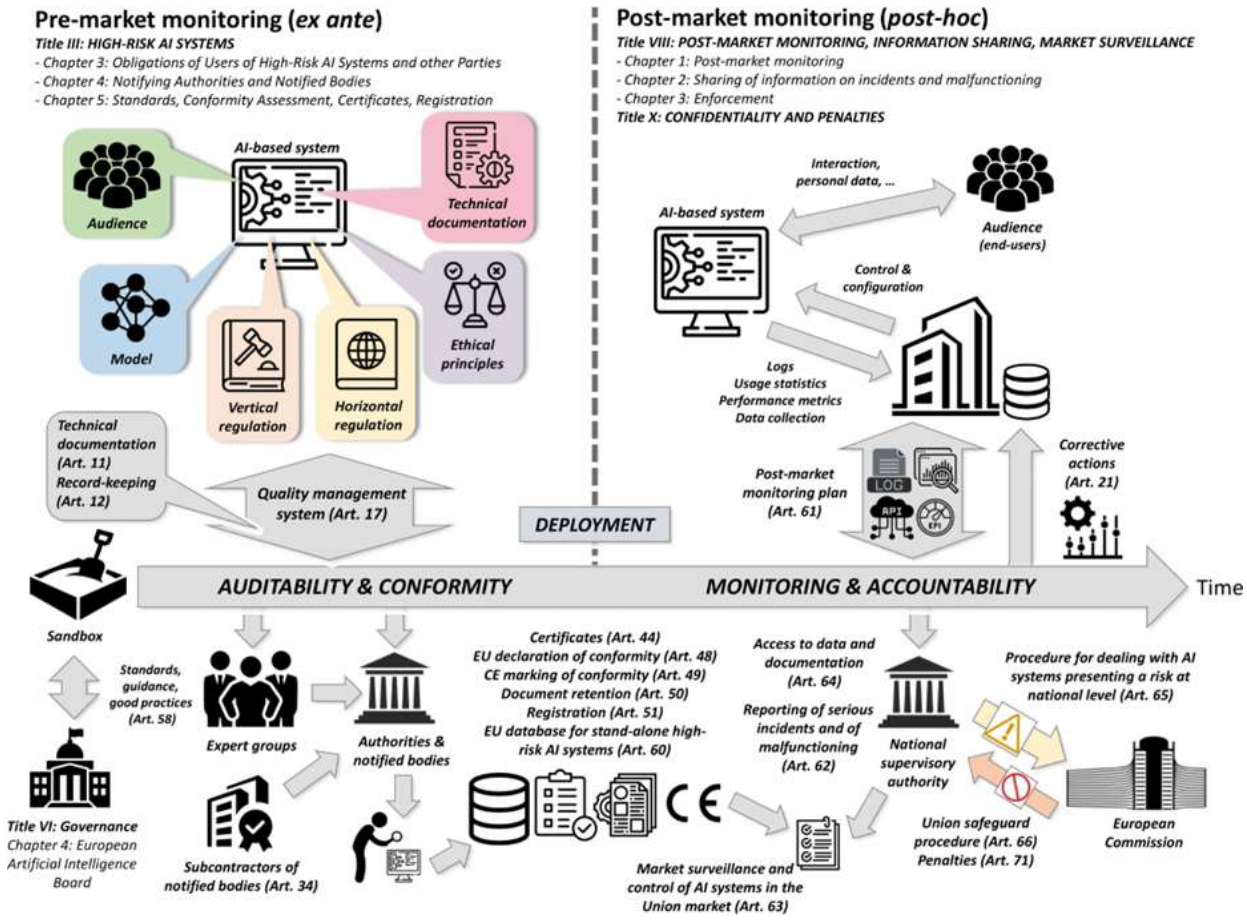

Figure 1. Diagram showing the role of auditability before (ex-ante) and accountability after (post-hoc) the AI-based system has been deployed in the market (Source: [Dia23]).

In parallel to the technical requirements, we have to pay attention to the regulation, with an approach based on **levels of risk.** In Europe, regulatory requirements in force for the deployment of *AI systems* are prescribed based on the risk of such systems to cause harm. Indeed, the AI Act agreed by the European Parliament, the Council of the European Union, and the European Commission, is foreseen to set a landmark piece of legislation governing the use of AI in Europe and regulating this technology based on the definition of different levels of risks: minimal, limited, high-risk and unacceptable risk. In these categories different requirements for trustworthy AI and levels of compliance are established [Dia23].

It is important to note that auditability refers to a property sought for the AI-based system, which may require transparency (e.g. explainability methods, traceability), measures to guarantee technical robustness, etc. Note that the auditability of a r*esponsible AI system* may not necessarily cover all requirements for *trustworthy AI*, but rather those foretold by ethics, regulation, specifications and protocol testing adapted to the application sector (i.e., vertical regulation).

We talk about risk levels, and we must also talk about ***high-risk scenarios***. The AI Act introduces the ***High-risk AI systems*** (HRAIs) as similar concept of *responsible AI systems* for high-risk scenarios, as systems that can have a significant impact on the life chances of a user (Art. 6); they create an adverse impact on people's safety or their fundamental rights. Eight types of systems fall into this category (that is, eight high-risk scenarios). These are subject to stringent obligations and must undergo conformity assessments before being put on the European market, e.g. systems for eligibility for public benefits or assistance, or law enforcement or access to education. They will always be high-risk when subject to third-party conformity assessment under that sectorial legislation.

A complete discussion on *Responsible AI systems* for a high-risk scenario leads us to stablish a set of auditability requirements and metrics to design the mentioned methodologies. Key attributes such as robustness, explainability, transparency and traceability, sustainability, fairness

are essential among others. See [Giu24] and [Fer23] for an initial analysis in two different contexts, financial services and autonomous driving domain respectively. This is an area that requires a great deal of attention and is a great challenge to establish compliance requirements and metrics, and tailored to each high-risk scenario.

We should delve into another essential aspect for responsible AI systems, safe AI. ***AI safety*** 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, including monitoring systems, adversarial robustness, detecting malicious use, attacks and backdoors, … Beyond AI research, it involves developing norms and policies that promote safety [Hen21].

Last but not least in the holistic view, any analysis must be accompanied by another critical aspect dedicated to ethics and all its social implications. It is necessary to consider the social acceptance or economic and legal implications, thus analysing the **ELSEC aspects of AI-based systems** (ethical, legal, socioeconomic and cultural).

Finally, we would like to conclude by stressing that safe and trustworthy AI is a critical area to meet upcoming regulations, the necessary auditability metrics for their analysis and compliance, address ethical issues, manage risk analysis in human-AI system interaction, and ensure the technical soundness of *responsible AI systems*.

This is the beginning of a fascinating path that enables the development of technology for the development of *responsible AI systems*. The goal of a *responsible AI system* is to employ AI in a safe, reliable and ethical manner. The journey is just beginning and in the next few years we will have auditable AI systems and auditability methodologies in all the necessary high-risk scenarios.

Acknowledgement: I would like to thank the co-authors of the paper [Dia23] and the members of the Spanish STAIRS (Safe and trustworthy AI) network proposal for the enriching discussions. These have allowed me to come up with the present lecture and a global view of the topic.

References

[AIA21] European Commission, 2021. Artificial Intelligence Act (Laying down harmonised rules on artificial intelligence and amending certain union legislative acts), Accessible at: https://artificialintelligenceact.eu/the-act/

[Dia23] 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.

[Bak21] Baker-Brunnbauer, J., (2021). TAII Framework for Trustworthy AI Systems. Robonomics: The Journal of the Automated Economy, 2, 17.

[Fer23] Fernández-Llorca, Gómez, E (2023). Trustworthy Artificial Intelligence Requirements in the Autonomous Driving Domain. *Computer*, vol. 56, no. 2, pp. 29-39.

[Giu24] Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial Intelligence risk measurement. *Expert Systems with Applications*, p.121220.

[Hen22] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916 (Version v5).*

[Kau22] Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2), 1-38.

[Li23] Li, B., Qi, P., Liu, B., et al. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1-46.

[Sin23] Singh, A. M., & Singh, M. P. (2023). Wasabi: A conceptual model for trustworthy artificial intelligence. Computer, 56(2), 20-28.