Responsible AI through a Software Engineering lens @ Serlab

Maria Teresa Baldassarre[1], Vita Santa Barletta[1], Danilo Caivano[1], Domenico Gigante[1], and Azzurra Ragone[1]

[1]Dipartimento di Informatica, Università degli Studi di Bari "A. Moro", Campus E. Quagliarello, Bari, 70125, Italy

Abstract

In this paper we summarize the research project currently conducted by the Software Engineering Research LABoratory (SERLAB) at Dipartimento di Informatica, Università degli Studi di Bari "A. Moro" on the topic of Responsible Artificial Intelligence (RAI).

Keywords

Trustworthy AI, Responsible AI, Software Engineering, Framework

1. Introduction

Artificial Intelligence (AI) is a revolution that is reshaping science and society as a whole [1]. While AI-related technologies are changing how data is processed and analyzed [2], autonomous and semi-autonomous decision systems are being used more frequently in several industries, such as healthcare, automotive, banking, and manufacturing, just to cite a few [3]. Given AI revolutionary potential and wide-ranging social influence, there has been a lot of discussion regarding the values and principles that should lead its development and application [4][5]. Recent scientific research and media attention have been focused on concerns that AI may endanger the jobs of human workers [6], be abused by malicious actors [7], avoid responsibility, or accidentally spread bias and as so, erode fairness [8].

In this particular context, the concept of Responsible Artificial Intelligence (RAI) started emerging. Cheng et al. [9] provide the following definition: "*intelligent algorithms that prioritize the needs of all stakeholders as the highest priority, especially the minoritized and disadvantaged users, in order to make trustworthy decisions. These obligations include protecting and informing users, preventing and mitigating negative impacts, and maximizing the long-term beneficial impact. (Socially) Responsible AI Algorithms constantly receive feedback from users to continually accomplish the expected social values*".

Other documents use the term *Trustworthy* instead of Responsible. This is the case of the resources published by OECD AI Policy Observatory (OECD.AI[1]). Since these terms can be treated as synonyms, for the sake of simplicity, in this document we will proceed using only Responsible Artificial Intelligence (RAI).

Several public and private organizations have responded to these societal fears by developing different kinds of resources: ethical requirements, principles, guidelines, best practices, tools, and frameworks.

In this work we briefly summarise our research activities intended to support, by shedding lights on problems and providing possible solutions, the realisation of a more responsible AI world.

The remainder of the paper is organized as follows: Section 2 defines some background definitions useful to set the stage; Section 3 highlights the research problem we are studying; Section 4 describes our vision and the research questions we are trying to answer; Section 5 summarizes our most important findings to date and, finally, our planned future works are drawn in Section 6.

2. Background

In this section we provide some preliminary definitions to better understand the concepts that guide our research.

[1]https://oecd.ai/en/

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

Corresponding author.

mariateresa.baldassarre@uniba.it (M. T. Baldassarre); vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); domenico.gigante1@uniba.it (D. Gigante); azzurra.ragone@uniba.it (A. Ragone)

0000-0001-8589-2850 (M. T. Baldassarre); 0000-0002-0163-6786 (V. S. Barletta); 0000-0001-5719-7447 (D. Caivano); 0000-0003-3589-6970 (D. Gigante); 0000-0002-3537-7663 (A. Ragone)

2.1. Responsible AI Principles

National and international organizations have created ad-hoc expert groups on AI to address the risks connected with the development of AI, frequently with the task of generating policy documents. These organizations include, among others, the High-Level Expert Group on Artificial Intelligence established by the European Commission[2], the UNESCO Ad Hoc Expert Group (AHEG) for the Recommendation on the Ethics of Artificial Intelligence[3], the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore[4], the NASA Artificial Intelligence Group[5] and the UK AI Council[6], just to cite a few.

These committees have been appointed to produce reports and guidelines about Responsible AI. Similar initiatives are being made in the commercial sector, particularly by businesses that depend on AI. Businesses like Sony[7] and Meta[8] made their AI policies and principles available to the public. At the same time, professional organizations and no-profit groups like UNI Global Union[9] and the Internet Society[10] have all released statements and recommendations.

The significant efforts of such an ample group of stakeholders to develop RAI principles and policies not only show the need for ethical guidance but also point out their keen interest in reshaping AI ethics to suit their individual priorities [10]. Notably, the private sector's participation in the field of AI ethics has been questioned since it may be using high-level soft policy as a portmanteau to either make a social issue technical [10] or avoid regulation altogether [11, 12].

However, many research works highlighted how these proposals often diverged, giving different definitions, resulting in the problem known as *principle proliferation* [13]. Consequently, several in-depth investigations have been conducted, such as the one by Jobin et al. [12], who found a global convergence around five ethical principles: *transparency*, *justice and fairness*, *non-maleficence*, *responsibility*, and *privacy*. Jobin et al. [12] stated that no one of these ethical principles is present in all the documents they reviewed; however, these five principles are mentioned in more than half of all the sources reviewed. Moreover, further in-depth thematic analysis revealed notable semantic and conceptual divergences in interpreting these principles and in the particular recommendations or areas of concern drawn from each of them.

2.2. Chosen AI principles definitions

As highlighted in Section 2.1, there are a lot of uncertainties and nuances around the definition of the principles that mainly characterize RAI, as well as, about the definition of RAI itself.

In our research, to address the problem of *principle proliferation*, we have decided to focus on a specific subset of principles, in particular, the four principles identified by Jobin et al. [12] with the exclusion of *responsibility* as this concept is rarely defined in a clear manner.

Moreover, to give an authoritative and clear definition for each principle, we decided to consider the ones provided by the High-Level Expert Group on Artificial Intelligence established by the European Commission[11] in their *Ethics guidelines for trustworthy AI* [14].

The resulting chosen principles are *transparency* (often known as *explainability*), *justice and fairness*, *non-maleficence* (often known as *security*) and *privacy*.

2.3. Frameworks

Another key element in our research are frameworks.

The concept of *framework* is far well-known in the Software Engineering (SE) field. Already in 1997, Johnson et al. [15] referred to frameworks as "*an object-oriented reuse technique*" or "*the skeleton of an application that can be customized by an application developer*". These are not conflicting definitions; the first describes the structure of a framework while the second describes its purpose.

Shifting the focus from SE to a more general context, frameworks are a form of design reuse.

Frameworks can be considered a collection of suggestions, guidelines and tools to be followed in order to create a product compliant with a defined standard.

[2] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[3] https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

[4] https://www.cms-holbornasia.law/en/sgh/publication/singapore-to-form-advisory-council-for-ethical-use-of-ai

[5] https://ai.jpl.nasa.gov/

[6] https://www.gov.uk/government/groups/ai-council

[7] https://www.sony.com/en/SonyInfo/sony_ai/responsible_ai.html

[8] https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

[9] http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

[10] https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/

[11] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Of fundamental importance in our research are frameworks that implement the above-mentioned RAI ethical principles.

2.4. Software Development LifeCycle (SDLC)

The concept of *Software Development Life Cycle* (SDLC) represents a fundamental piece of knowledge in the field of Software Engineering (SE). Boehm et al. [16] in 1976 already were talking about common activities involved in the production of a software system. Subsequently, these activities have been standardised in 2017 into the ISO/IEC 12207[12]. These standardized activities which compose the SDLC are:

- *Requirement Gathering and Analysis.*
- *Design.*
- *Implementation or Coding.*
- *Testing.*
- *Deployment.*

According to the development methodology used (e.g. Agile[13] or Waterfall[14]), these activities may be done sequentially or iteratively.

3. Problem statement

Neglecting RAI precautions may lead to several threats for the end users. In the following, we give an overview of the four ethical principles we choose to address in our research.

3.1. Justice and Fairness

AI models can amplify existing bias coded in data or introduce new forms of bias [17], resulting in unfair decisions in legal or ethical sense. In particular, AI-based systems may produce decisions or have impacts that are discriminatory or unfair, and this is especially true when AI is deployed in complex socio-technical systems.

Note that cases like these do not have to be intentional on the part of the people who designed/developed these systems. Instead, issues like these may arise, for example, due to the datasets or algorithms used to develop the systems.

An example of "justice and fairness issue" in an AI system is a famous smart algorithm guiding care for tens of millions of people which in 2019 was found to be biased against dark-skinned patients in New Jersey, USA; the effects were it was assigning dark-skinned patients lower scores than white patients with the same medical conditions[15]. This could have caused wrong medical prescriptions.

3.2. Non-maleficence (or Security)

Since AI models learn from data, there exists the possibility a malicious actor can provide manipulated training samples which force the model to learn from a distorted reality.

This threat is known as *adversarial machine learning*, which poses great challenges to deep learning models. This threat is particularly harmful in safety-critical scenarios, such as self-driving, where the vision system must be robust to ad-hoc crafted external perturbations. An adversarial example is a malicious input typically created by applying a small but intentional perturbation, such that the attacked AI model misclassifies it with high confidence.

Different specific instances of this threat exist: Adversarial Examples, Data Poisoning, Model Evasion, Trojan (or Backdoor) and Model stealing (or Model Extraction).

The victim of such an attack was the chatbot "Tay", developed by Microsoft in 2016, which was shut down and closed a few hours after its release time because was attacked and forced to post offensive tweets against users[16].

3.3. Transparency (or Explainability)

In recent years, sophisticated AI models are being applied in real contexts, assisting humans in the most disparate domains. However, the increase of their performance has been accompanied by an increase of complexity at a level that no one, neither their designers, can interpret the inner workings leading to decisions.

Many researchers started focusing on the urgent open challenge of how to construct meaningful explanations for opaque (i.e. systems whose functioning logic is not comprehensible by a human) AI systems in the context of AI-based decision-making, aiming at empowering individuals against undesired effects of automated decision-making, implementing the "right of explanation", helping people make better decisions preserving (and expand) human autonomy.

[12] https://www.iso.org/standard/63712.html
[13] https://www.sciencedirect.com/topics/computer-science/agile-methodology
[14] https://www.sciencedirect.com/topics/computer-science/waterfall-methodology
[15] https://www.nature.com/articles/d41586-019-03228-6
[16] https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

This research branch is known as eXplainable AI (XAI).

This is the only principle which does not pose direct harm to the final human users of the model. Anyway, an understandable AI algorithm would instil confidence in its users and help its owners in understanding and debugging unexpected decisions.

3.4. Privacy

Privacy is currently one of the first human rights that has been considered in legal frameworks and regulations, e.g. the European GDPR [18]. As a consequence, even if a lot of work has been done in scientific literature, there is still the need to investigate new methodologies and approaches to (a) formally define and automatically detect privacy risks raised by AI systems handling different kinds of personal data; (b) design data anonymization algorithms that are sufficiently robust to sophisticated attacks but computationally feasible; (c) design AI algorithms or plugins able to enforce privacy by design constraints; (d) investigate existing measures (or create new ones) to evaluate the privacy risk of novel or unusual kinds of data.

The privacy attacks the literature already conducted against AI models are Membership Inference and Model Inversion, aimed to reconstruct the data on which the model was trained or the model itself.

Another aspect to consider is the possible data leakage which can be caused by the interoperability of AI models with classical software. Just a few days ago, celebrity ChatGPT has an issue with the titles given to its users' chats[17]: users' conversations were randomly exposed to other users without consent, which may be a violation of GDPR regulation by OpenAI (the company behind ChatGPT). The root cause was a "major issue" with a third-party open-source library, according to the company, which has subsequently been resolved. This incident, added to other missing protections, cost ChatGPT a ban in Italy from the Italian data-protection authority[18].

4. Goal and Research Questions

In this section, we highlight the Goal of our research, the conceptual roadmap we follow, the Research Questions (RQs), and the Actions (A) we tackle to answer the RQs.

[17] https://techmonitor.ai/technology/ai-and-automation/chatgpt-bug-openai-gdpr

[18] https://www.bbc.com/news/technology-65139406

Goal

Provide AI practitioners, both technical and non-technical stakeholders, with guidelines, best practices, and tools to support and guide the development of Responsible AI applications in all the Software Development Lifecycle (SDLC).

Starting from this goal, we planned a conceptual roadmap made of sequential and correlated steps: realize a framework prototype that helps different kinds of stakeholders to address Responsible AI issues.

This roadmap starts with the study of the current literature in the field of Responsible AI, with the goal of understanding what has been done so far and collecting the discovered gaps and needs.

Since the literature may occasionally miss particular problems encountered by the AI practitioners, the roadmap includes a study on the field directly with the practitioners, to understand their real needs and validate the results obtained from the literature study.

Finally, the roadmap envisages the development of a **framework prototype to guide different stakeholders — both with a technical or non-technical background — in the development of AI applications**.

One peculiarity of this prototype is that at its best it **should cover all different phases and activities of the SDLC**: we consider this a mandatory requirement because, while realizing a product or a service, several stakeholders collaborate to achieve their common final goal — i.e. the product or service — but each one has different skills and provides a different point of view for the product (or service).

We expanded this overall goal in a few macro research questions:

- **RQ1**: What is the state of the practice and the correlated literature to approach the Responsible AI development?
- **RQ2**: What do the practitioners think about Responsible AI? What are their perceived gaps?
- **RQ3**: Is it possible to realize a framework able to support different kinds of stakeholders in implementing Responsible AI?

In order to address each of these questions, in the following there are the **Actions (A)** we intend to perform:

- **A1**: Perform a rapid review to collect all published resources — such as frameworks and tools — related to Responsible AI.

- **A2**: Spread a survey and conduct focus groups with AI practitioners, both from academia and industry, to understand if they agree with the problems that emerged from the literature and discover new specific gaps and necessities.
- **A3**: Define a framework prototype to support AI practitioners in the development of AI applications. This framework should include both guidelines and tools related to RAI.

5. Current practice gaps

To answer RQ1, in [19] we investigated the state of the literature and of the practice by doing a rapid review of most of the frameworks proposed by both public and private entities to address Responsible AI issues. An overview of the findings is presented below; the interested reader can find all the details in [19].

In November 2022 we consulted various search engines, to collect both white- and grey-literature resources. The research lasted one month and ended up with 148 unique resources (without duplicates).

All the retrieved sources were classified w.r.t. the type of proposing institution (COMPANIES, UNIVERSITIES, and NO-PROFIT ORG / COMMUNITIES / PUBLIC ENTITIES (NPG/COMM/PE)) and according to their type (Principle (P), Guideline (G), Tool (T) or Other (O)).

First of all, our analysis highlighted that most of the filtered frameworks are proposed by NPG/-COMM/PE (50.7%). Regarding the type, we can say that there is a worrying lack of tools: most of the frameworks are just Principles or Guidelines.

A positive trend is that the majority of the frameworks address all four principles presented in Sec. 2.2, even if sometimes in a "partial" way: this reveals an even greater lack of consensus and standardization about which are the best practices to follow to be compliant with the RAI values.

Nevertheless, some frameworks neglect one, two, or even three of these principles.

Anyway, a negative trend is that very few frameworks encompass all the SDLC phases, thus not providing practical support to practitioners who want to develop, test, and deploy RAI applications; most frameworks focus only on the initial phases of the SDLC, and in particular on *Requirements elicitation*.

Finally, another negative trend is that in most cases there is not a practical tool complementing the theoretical frameworks; this is true regardless of the type of entity releasing the tool.

All these findings are also supported by the common worries publicly raised by some experts on AI, who stated "*AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research*"[19]. To summarize, right now does not exist any comprehensive framework whose knowledge can be navigated and exploited by different kinds of stakeholders (technical and non-technical ones), which can simplify and speed up the adoption of RAI practices.

6. Challenges and Future Work

As already mentioned in Sec. 4, what we discovered while analyzing the literature may be different from what practitioners actually need. This leads now to continue our research by validating in the practical field our previous findings.

Our next step consists in spreading a survey among AI experts (both from industry and academia) to collect as much structured data as possible, in order to derive an initial preview of the actual practical gaps in the state of the practice. By analyzing this data, we aim to extract the key points requiring a deeper investigation. Then we intend to eviscerate these key points by conducting focus groups in which we ask the practitioners if they agree regarding the gaps that emerged from literature on Responsible AI. We will also ask for their points of view, possibly by showing real-world cases in which they would have had suggestions regarding Responsible AI. This formalized data will enable us to answer RQ2.

Then, once the practitioners' needs will be clearly elicited, we will be one step away from our main goal to develop a comprehensive framework able to provide support to multiple different stakeholders while addressing Responsibility issues in AI across the entire SDLC. We plan to include information at different abstraction levels and coming from different knowledge domains (e.g., legal laws as well as best practices for software development). A possible solution may be the formalization of a knowledge base, but we must be careful to keep knowledge in a structured and organized form, in order to facilitate its query.

Before obtaining a useful and user-friendly framework, the underlying knowledge base must be validated on real-world AI systems, both open- and private- source, to identify possible gaps and apply the required refinements.

[19] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Finally, we plan to realize an automated and user-friendly interface which should support the various stakeholders and possibly automate repetitive tasks.

References

[1] Y. N. Harari, Reboot for the ai revolution, Nature 550 (2017) 324–327. URL: https://doi.org/10.1038/550324a. doi:10.1038/550324a.

[2] M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (2015) 255–260. doi:10.1126/science.aaa8415.

[3] G. Cornacchia, F. Narducci, A. Ragone, Improving the user experience and the trustworthiness of financial services, in: Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference, volume 12936 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 264–269. URL: https://doi.org/10.1007/978-3-030-85607-6_19. doi:10.1007/978-3-030-85607-6_19.

[4] E. Vayena, A. Blasimme, I. Cohen, Machine learning in medicine: Addressing ethical challenges, PLOS Medicine 15 (2018) e1002689. doi:10.1371/journal.pmed.1002689.

[5] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (2018). doi:10.1038/s41586-018-0637-6.

[6] n.d., Science must examine the future of work, Nature News 550 (2017) 301–302. URL: https://doi.org/10.1038/550301b. doi:10.1038/550301b.

[7] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. Allen, J. Steinhardt, C. Flynn, S. hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, D. Amodei, The malicious use of artificial intelligence: Forecasting, prevention, and mitigation (2018).

[8] J. Zou, L. Schiebinger, Ai can be sexist and racist — it's time to make it fair, Nature 559 (2018) 324–326.

[9] L. Cheng, K. R. Varshney, H. Liu, Socially responsible ai algorithms: Issues, purposes, and challenges, Journal of Artificial Intelligence Research 71 (2021) 1137–1181.

[10] D. Greene, A. L. Hoffmann, L. Stark, Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning, in: Hawaii International Conference on System Sciences, 2019.

[11] Wagner, B. in *Being Profiled: Cogitas Ergo Sum. 10 Years of 'Profiling the European Citizen'* (eds Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M.), Amsterdam University Press, 2018. URL: https://mediarep.org/handle/doc/14277. doi:10.1515/9789048550180.

[12] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. doi:10.1038/s42256-019-0088-2.

[13] L. Floridi, J. Cowls, A unified framework of five principles for ai in society, Issue 1 (2019).

[14] High-Level Expert Group on AI (AIHLEG), Ethics guidelines for trustworthy AI | Shaping Europe's digital future, 2018. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[15] R. E. Johnson, Frameworks = (components + patterns), Communications of the ACM 40 (1997) 39 – 42. doi:10.1145/262793.262799.

[16] B. W. Boehm, Software engineering, IEEE Transactions on Computers C-25 (1976) 1226 – 1241.

[17] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven ai systems – an introductory survey, 2020. arXiv:2001.09762.

[18] E. Parliament, of the Council of European Union, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), 2016.

[19] V. S. Barletta, D. Caivano, D. Gigante, A. Ragone, A rapid review of responsible ai frameworks: How to guide the development of ethical ai, The International Conference on Evaluation and Assessment in Software Engineering (EASE) 2023 (To appear).