POLARIS: A framework to guide the development of Trustworthy AI systems

Maria Teresa Baldassarre[1], Danilo Caivano[1], Domenico Gigante[1,2,*] and Azzurra Ragone[1,*]

[1] *University of Bari "A. Moro", Bari, Italy*
[2] *Ser&Practices Srl, Bari, Italy*

Abstract

In the ever-expanding landscape of Artificial Intelligence (AI), where innovation thrives and new products and services are continuously being delivered, ensuring that AI systems are designed and developed responsibly throughout their entire lifecycle is crucial. To this end, several AI ethics principles and guidelines have been issued to which AI systems should conform. Nevertheless, relying solely on high-level AI ethics principles is far from sufficient to ensure the responsible engineering of AI systems. In this field, AI professionals often navigate by sight. Indeed, while recommendations promoting Trustworthy AI (TAI) exist, they are often high-level statements difficult to translate into concrete implementation strategies. Currently, there is a significant gap between high-level AI ethics principles and low-level concrete practices for AI professionals. To address this challenge, in this discussion paper we describe the novel holistic framework for Trustworthy AI we developed — designed to bridge the gap between theory and practice. The framework builds up from the results of a systematic review of the state of the practice as well as a survey and think-aloud interviews with 34 AI practitioners. The framework, unlike most of the ones in the literature, is designed to provide actionable guidelines and tools to support different types of stakeholders throughout the entire Software Development Life Cycle (SDLC). Our goal is to empower AI professionals to confidently navigate the ethical dimensions of TAI through practical insights, ensuring that the vast potential of AI is exploited responsibly for the benefit of society as a whole.

Keywords

Artificial Intelligence, Software Engineering, Trustworthy AI, Knowledge Base, Framework

1. Introduction

In the dynamic realm of Artificial Intelligence (AI), marked by ceaseless innovation and rapid advancements, the ethical, societal, and operational implications of AI technologies have shifted to the forefront of discussions. As AI systems become deeply integrated into our daily lives, from healthcare [1] to finance [2], and influence critical decision-making

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.
mariateresa.baldassarre@uniba.it (M. T. Baldassarre); danilo.caivano@uniba.it (D. Caivano); d.gigante@serandp.com (D. Gigante); azzurra.ragone@uniba.it (A. Ragone)
0000-0001-8589-2850 (M. T. Baldassarre); 0000-0001-5719-7447 (D. Caivano); 0000-0003-3589-6970 (D. Gigante); 0000-0002-3537-7663 (A. Ragone)

processes, the responsible development and deployment of AI has transitioned from an academic discourse to an imperative in real-world applications.

In this complicated context, the concept of Trustworthy Artificial Intelligence (TAI) [3] has grown relevance. These risks are even more pronounced with the recent advent of Generative AI — e.g. ChatGPT — and how this impacts on various societal aspects [4].

Several public and private organizations have tried to address TAI by developing different kinds of resources, just to cite a few: ethical requirements [5], principles [6], guidelines [7], best practices [3], tools [8], and frameworks [9]. However, navigating the intricacies of TAI has become increasingly complex due to what we might refer to as "*principle proliferation*" [10]. This phenomenon encompasses the multitude of ethical principles that have been devised, each one providing a specific definition, but also contributing to a landscape that can overwhelm AI practitioners.

In response to the challenges posed by principle proliferation, our research follows the work of Jobin et al. [10] and focuses on four foundational pillars of TAI: **Privacy**, **Security**, **Fairness**, and **Explainability**. These pillars have been condensed in a practical, focused, and adaptable framework, called POLARIS [11]. Its aim is to guide AI practitioners and stakeholders in their quest to ensure the effective trustworthy development of AI-enabled systems across the entire software Development LifeCycle (SDLC).

The main contributions of this work — detailed in [11] — can be summarized as follows:

- A review of the **state of the practice** and identification of **practitioner needs** to understand existing practices, challenges, and what practitioners currently lack in developing Trustworthy AI applications.
- The proposal of a **novel framework** (POLARIS) that systematizes and organizes the knowledge found in different sources. The objective of POLARIS is to make this **knowledge easily accessible** to AI practitioners and to provide them with **actionable guidelines** that can be applied in every phase of the SDLC.

The paper is organized as follows: Section 2 outlines the results of the systematic review and the findings from the survey and interviews with AI professionals. Section 3 describes the POLARIS framework, its components and how to use it. Conclusions are drawn in Section 4.

2. State of the practice

This discussion paper builds on previous research [12], in which we conducted a comprehensive study of the state of the practice of existing TAI frameworks. More precisely, we investigated (i) the extent to which the analyzed frameworks addressed the principles mentioned in Section 1 and (ii) whether and to what extent these frameworks covered the stages of the Software Development Life Cycle (SDLC). Next, we carried out a comparative analysis among the identified frameworks with respect to characteristics such as best practices, guidelines, and tools, in order to assess if and how big a gap there is between the proposed high-level AI ethics principles and low-level operational practices for practitioners.

In our previous work [12] we analyzed **138 frameworks**, both from white-literature and grey-literature sources. The main findings are:

1) Most of the frameworks are proposed by *No-profit Organisations, Public Entities or Human Communities* (50.7%); followed by private Companies (31.9%) and then Universities (17.4%).

2) Most of the frameworks provide a set of TAI principles/values (46.1%); others include actionable Guidelines (29.6%), but very few also provide Tools (9.2%).

3) In the majority of cases, *the frameworks address all four TAI principles* (45.1%) even if there are frameworks that cover only one (15.5%) or two (15.5%) principles;

4) More than half of the frameworks (55.2%), provide support only for the *Requirements Elicitation* phase. While all the SDLC phases are covered only in 5.7% of the frameworks;

5) In more than 80% of the cases there is *no tool included in the framework* and when it is present, it is directed to *Non-technical* stakeholders (i.e. stakeholders who work in the first two phases of the SDLC — e.g. commercial agents, functional analysts, architecture designers, etc).

In summary, our analysis confirmed that most of the existing frameworks include high-level best practices, checklists, or self-assessment questions, most suitable for non-technical stakeholders, limitedly able to address technical stakeholder needs and close the gap between high-level principles definition and practical recommendations for AI practitioners covering all the SDLC phases. The findings from [12] have then guided our subsequent research aimed at investigating the needs of AI practitioners, their current practices and issues encountered in the design and implementation of trustworthy AI systems.

2.1. Identification of Practitioner needs

As a premise for the design and proposal of our framework, we conducted an exploratory survey [13] to collect practitioner insights and needs with respect to TAI principles. We used convenience sampling and recruited practitioners from companies in our network of collaborations. All participants were practitioners with experience in developing AI-enabled systems who had addressed, to some extent, TAI in their projects. We contacted a total of 45 professionals, of which 34 completed the survey. These participants represented a diverse spectrum, ranging from small-medium companies (55.9%), to large companies with more than 1000 employees (44.1%).

Apart from the demographics, the survey is organized into three main parts, each focusing on the collection of specific data pertaining: a) existing practices, b) identification of challenges, c) discovery of unmet needs. In the following, for each part, we provide a brief explanation and highlight the main results.

A) Exploration of Existing Practices. The first part of the survey investigated the operational procedures and methodologies employed by practitioners in the context of implementing TAI. This exploration aimed to provide a detailed insight into the real-world practices and strategies adopted.

Results. First, we observed that the TAI principle most frequently addressed by participants is *Privacy* (58.8%), and most of the participants address at least one TAI principle

during *Design* (64.7%) and *Development* (47.1%) SDLC phases. On the contrary, very few participants declared to address at least one TAI principle during the *Test* (29.4%) and *Deploy* (20.6%) phases. This may highlight the need for more support, in terms of tools and guidelines in the last phases of the SDLC. Moreover, when participants faced issues related to TAI, in half of the cases they did not even try to address or solve them (50%). Probably because they did not know how to or they simply considered them not worth solving. This is a point that deserves more investigation. Only 35% of the participants declared to have directly addressed TAI issues, while a small percentage stated that the issue resolution was demanded to a third party (15%).

B) Identification of Challenges. This section of the survey explored the challenges and obstacles encountered by professionals while trying to integrate TAI into their systems. By identifying these issues, we aimed to shed light on critical areas where AI professionals may need more support.

Results. In cases where the respondents tried to address/fix TAI issues, we found that the most voted impediments are: (i) "*the issue solution required too much time to be implemented*" (58.3%) and (ii) "*the issue solution was likely to decrease the performance of the system (e.g., decreasing accuracy)*" (50%). On the other hand, none of the participants answered: "*no one had idea on how to solve the issue*", which is a positive result since it indicates that practitioners are conscious of untrustworthiness problems and are able to hypothesize solutions. Among the comments, one participant mentioned "*[scarce] data availability*" as an impediment.

C) Discovery of Unmet Needs. The third part of the survey reveals the presence of unaddressed needs within the practical landscape of TAI. Specifically, we uncovered a range of requirements that have so far received limited attention within the existing literature.

Results. Regarding the prevention of trustworthiness issues in AI, the participants rated as the most valuable tool able to "*[...] generate an explanation of a model after its creation [...]*" (with 82% of positive answers) while they rated as least useful (i) a tool to help "*deciding how much data you need for particular subgroups/subpopulations*" and (ii) a tool to "*generate possible adversarial/malicious data points to test to use in testing the system*" (both with 19% of negative answers).

On the other side, to address untrustworthiness in AI, the participants rated as the most valuable (i) "*best practices that can actively guide your team through the model's SDLC*" (92% positive answers), (ii) a tool able to "*[...] help [...] monitoring the AI model after its release to the public*" (91% positive answers), and "*a knowledge book in which are mapped trustworthiness problems and [...] solutions*" (70% positive answers). On the other hand, they rated as least useful (i) a tool "*[...] to help your team doing an ex-post TAI audit*" (18% of negative answers) and (ii) a tool able to *[...] help your team deciding which AI model best respects the TAI principles [...]* (17% of negative answers).

Overall, these results confirmed the findings of our previous work [12]: a significant majority of the respondents expressed the need for *comprehensive knowledge bases* and *pragmatic guidelines* offering insights and recommendations for the seamless implementation of trustworthy AI system throughout the entire SDLC. Furthermore, they also highlighted the lack of *tools* supporting them in the last stages of SDLC.

3. The POLARIS Framework

In response to the challenges and issues highlighted by the research results, with the intent to fill the gap between theory and practice and to address stakeholder needs and shortcomings (Section 2), we have developed a framework: POLARIS [11]POLARIS.

Indeed, POLARIS has been designed to provide actionable guidelines and tools in order to support stakeholders in addressing TAI principles throughout the entire Software Development Life Cycle (SDLC). POLARIS provides a significant amount of information, organized and linked into a comprehensive knowledge base that is designed to be expandable, with the possibility to easily add new knowledge.

In Section 3.1 we explain how we built the POLARIS knowledge base while in Section 3.2 we describe how to navigate it.

3.1. Defining POLARIS Knowledge Base

In this section, we describe how we assembled the POLARIS Knowledge Base and the selection process used to choose the different knowledge sources representing the foundation of POLARIS.

We started from the frameworks analyzed in [12], we complemented our analysis with the results obtained from the survey, and then we identified among the existing knowledge sources (i.e. frameworks) those that met both of the following criteria:

1. Have actionable guidelines (and not only a simple high-level principles list)
2. Address all SDLC phases.

Regarding the last criterion, since SDLC phases do not always map with the activities required to develop an AI-enabled system, we have integrated each SDLC phase with AI-enabled activities established by Zhengxin et al. [14].

After this first selection phase, we identified only three knowledge sources that meet both criteria (1) and (2). Then, we mapped each identified knowledge source to the corresponding TAI principle. For **Explainability** we selected Jin et al. - EUCA: the Explainable AI Framework [15], for **Fairness** we chose Amsterdam Intelligence - The Fairness Handbook [16]. For both **Privacy** and **Security** we selected ENISA - Securing Machine Learning Algorithms [17].

Then, we refined this first selection by adding more knowledge sources that could complement the information provided by the primary ones initially selected. We started by selecting the frameworks that met *at least one* of the following criteria:

1. Have actionable guidelines (and not only a simple high-level principles list)
2. Address all SDLC phases.

We retrieved 10 additional knowledge sources that met at least one of the previous criteria. The table with all the 10 knowledge sources identified can be found in the online appendix [18]. Then, we performed a comparative analysis between each primary knowledge source already selected ([15], [16], [17]) and the new ones retrieved in this second iteration.

The results of the comparative analysis brought us to select four additional knowledge sources that could complement and expand the information provided by the first ones selected.

The additional frameworks selected were ICO's "*Guidance on AI and data protection*" [19], Tensorflow's "*Responsible AI in your ML workflow*" [20], the guidelines in Microsoft's "*Threat Modeling AI/ML Systems and Dependencies*" [21] and CSIRO's "*Responsible AI Pattern Catalogue*" [22]. Therefore, we used these additional knowledge sources to further extend the information provided by the primary ones. We ended up selecting **7 knowledge sources**. In the online appendix [18] there is the mapping between each TAI principle and the corresponding knowledge sources covering that principle.

3.2. Navigating POLARIS Knowledge Base

Having defined the POLARIS knowledge base, in this section we focus on how to navigate it. The goal of the proposed framework is to support stakeholders throughout the SDLC by suggesting concrete implementation strategies able to support and guide them in the development of TAI applications.

When applying the framework, the users will ultimately receive an *Action* to implement, that is, an actionable guideline that a stakeholder should consider and, if possible, implement while developing the AI-enabled software system to ensure compliance with the four TAI principles. The user can also choose to filter and apply only a subset of the suggested guidelines.

As of now, the first version of POLARIS has been structured as a filterable Excel array of sheets. There are **four** main **knowledge components**, one per each principle: (i) *Privacy*; (ii) *Security*; (iii) *Fairness* and (iv) *Explainability*.

In proposing the structure of each Excel sheet, we were inspired by the ENISA framework [17] and then customized it according to our needs.

The two sheets that contain the knowledge for "**Privacy**" and "**Security**" are composed of the following six columns (Fig. 1 shows an excerpt of the security component).

Figure 1: Excerpt of the Security component navigation of the POLARIS framework.

SDLC Phase	Threat	Sub-Threat	Description	Vulnerability (consequence)	Action
Design	Poisoning	Label modification	An attack in which the attacker corrupts the labels of training data. This sub-threat is specific to Supervised Learning.	Use of unreliable sources to label data	**(Technical) Ensure reliable sources are used**: ML is a field in which the use of open-source elements is widespread (e.g., data for training, including labeled ones, models). The trust level of the different sources used should be assessed to prevent using compromise ones. For example: the project wants to use labeled images from a public library. Are the contributors sufficiently trusted to have confidence in the contained images or the quality of their labelling?
Development	Failure or malfunction of ML application	Denial of service due to inconsistent data or a sponge example	ML algorithms usually consider input data in a defined format to make their predictions. Thus, a denial of service could be caused by input data whose format is inappropriate. It may also happen that a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service.	Use of uncontrolled data	**(Techincal) Control all data used by the ML model** Data must be checked to ensure they will suit the model and limit the ingestion of malicious data: - Evaluate the trust level of the sources to check it's appropriate in the context of the application - Protect their integrity along the whole data supply chain - Their format and consistence are verified - Their content is checked for anomalies, automatically or manually (e.g. selective human control) - In the case of labeled data, the issuer of the label is trusted.

(1) SDLC Phase. The SDLC phase that the *Action* column applies to.

(2) Threat. Contains the list of threats, i.e. possible attacks that can be conducted against an AI-enabled system. Examples are *Evasion* and *Poisoning* attacks.

(3) Sub-Threat. In some specific cases, a threat can have a specific declination in a sub-characteristic. For example, the *Poisoning* attack can be declined in *Targeted Data Poisoning* and *Indiscriminate Data Poisoning*.

(4) Description. A textual description of the (Sub)Threat, which helps the stakeholder obtain coarse-grained details about the threat and understand the attacker's objective.

(5) Vulnerability (consequence). This is the immediate consequence of having a model vulnerable to a specific threat.

(6) Action. The corresponding action, or decision, that should be adopted to address a specific threat, based on the SDLC phase and threat selected, keeping in mind the vulnerability.

For example, a developer in the *Design* SDLC phase who is trying to address the vulnerabilities associated with *Poisoning* threat, may consult POLARIS and access the *Security* Excel sheet, select the *Poisoning* threat — and corresponding sub-threat, i.e. *Label modification* —, and obtain a description of the vulnerability associated to the (sub)threat and the action to take in order to mitigate the vulnerability, i.e. *ensure that reliable sources are used* (Fig. 1).

The sheet that contains the knowledge for "**Fairness**" is composed of 5 columns, all of the above, except for *Vulnerability (consequence)* column which has been removed as the concept of vulnerability in the context of fairness does not apply.

The sheet containing the knowledge for "**Explainability**" has a different set of columns (see Fig. 2), because there are no real threats associated with the lack of explainability. However, having a system that is not explainable, will lead users to use it with some reluctance because of its opacity in making decisions, as it is not possible to derive any clear logical relationship between the internal configuration and their external behaviour, except for a few specific cases (e.g. decision trees) [23]. For Explainability, the columns are the following:

Figure 2: Excerpt of the Explainability component navigation of the POLARIS framework

SDLC Phase	Data type	Local/Global Explanation	Explanation Goal	Action
RE	General	Both	Start considering Explainability from Req. Elicitation	Elicit also explainability requirements; examples are: - Unexpected Prediction: Disagreement with AI: declare the required behaviour in case the AI prediction is unexpected, and/or users disagree with AI's prediction - Expected prediction: declare the required behaviour in case AI's prediction aligns with users' expectations - Differentiate similar instances: due to the consequences of wrong decisions, users sometimes need to discern similar instances or outcomes. For example, a doctor differentiates whether the diagnosis is a benign or malignant tumor - Learn from AI: users need to gain knowledge, improve their problem-solving skills, and discover new knowledge - Improve the predicted outcome: users seek causal factors to control and improve the predicted outcome - Communicate with stakeholders: many critical decision-making processes involve multiple stakeholders, and users need to discuss the decision with them - Generate reports: users need to utilize the explanations to perform particular tasks such as report production. For example, a radiologist generates a medical report on a patient's X-ray image
Design	General	Both	Have a clear idea about the desired explanation form	Consider the design of the explanation design: how the final UI should be composed and how to present the information

(1) SDLC Phase. The SDLC phase the *Action* column relates to.

(2) Data Type. The type of data used by the AI algorithm for which the action/guideline applies. Examples are *Tabular* data or *Image*. When the action applies to all algorithms,

regardless of the type of data, the tag *General* is used.

(3) Local/Global Explanation. This column describes the type of explanation that can be obtained by implementing the action. At the moment, the possible values are *Global* and *Local* [24].

(4) Explanation Goal. This is the goal that can be achieved if the action/guideline gets implemented. Examples are: *to validate the algorithm outcome* and *to reveal bias.*

(5) Action. The corresponding action, or decision, that should be taken to reach the selected explanation goal. We point out that for each <data, explanation type> pair there is at least a corresponding row in the framework.

For example, a user who is in the *Requirement Elicitation* SDLC phase and needs to enquire on all the possible explanation approaches to explain the output of an algorithm, could access the *Explainability* Excel sheet and select the *General* data type and retrieve a set of actions that pertain explainability requirements i.e. *elicit explainability requirements* (Fig. 2).

When navigating POLARIS, each stakeholder can use different filters and subfilters, based on specific needs, as for instance: *Knowledge Component* (i.e. TAI principle) to address, *Threat* (or *Sub-Threat*), *Vulnerability*, *SDLC Phase*, *Data type*, and *Local/Global Explanation.* One of the most significant filters is *SDLC phase*, which makes POLARIS flexible and allows stakeholders to use it either on ongoing/closed projects — where it is possible to address, for example, only the deployment or monitoring phase — or at the early stage of a project, since in the latter case it can cover all SDLC phases.

4. Conclusion

In this work, we described POLARIS, the framework we designed to fill the gaps highlighted in the review of the state of the practice and to provide AI practitioners with actionable guidelines specific to each phase of the SDLC.

POLARIS has four pillars (or *components*), which are Explainability, Fairness, Security, and Privacy. These principles have been chosen as they are the most recurrent TAI principles found in the current literature [10]. Each component provides practical guidelines and tools to support different kinds of stakeholders across the entire SDLC.

Its added value is that it provides knowledge already freely accessible online but in an organized and systematized way.

As detailed in [11], we identified several improvements. From a usability point of view, we are planning to (i) provide a more usable UI, like the one of the VIS-Prise tool [25]. Then, we plan to (ii) validate POLARIS on a growing number of case studies.

Moreover, if further validations confirm us the stakeholders are interested and plan to use POLARIS, in the next versions we plan to integrate more TAI principles.

POLARIS is a preliminary attempt to organize and make knowledge on TAI principles easily accessible and available to different kinds of stakeholders. It is a pioneering prototype whose goal is to make AI professionals, policymakers, and stakeholders able to navigate the ethical dimensions of TAI with confidence, ensuring that the vast potential of AI is harnessed responsibly for the benefit of society.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. M. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118. URL: https://api.semanticscholar.org/CorpusID:3767412.

[2] 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.

[3] 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.

[4] M. T. Baldassarre, D. Caivano, B. Fernandez Nieto, D. Gigante, A. Ragone, The social impact of generative ai: An analysis on chatgpt, in: Proceedings of the 2023 ACM Conference on Information Technology for Social Good, GoodIT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 363–373. URL: https://doi.org/10.1145/3582515.3609555. doi:10.1145/3582515.3609555.

[5] European Union, AI Act, 2023. URL: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[6] UNI Global Union, Top 10 Principles for Ethical AI, 2019. URL: https://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/.

[7] The Public Voice, Universal Guidelines for Artificial Intelligence, 2019. URL: https://thepublicvoice.org/ai-universal-guidelines/.

[8] Google, Tools & Platforms, 2019. URL: https://pair.withgoogle.com/tools/.

[9] NIST, AI Risk Management Framework, 2019. URL: https://www.nist.gov/itl/ai-risk-management-framework.

[10] 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.

[11] M. T. Baldassarre, D. Gigante, M. Kalinowski, A. Ragone, Polaris: A framework to guide the development of trustworthy ai systems, 2024. arXiv:2402.05340.

[12] V. S. Barletta, D. Caivano, D. Gigante, A. Ragone, A rapid review of responsible ai frameworks: How to guide the development of ethical ai, in: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 358–367. URL: https://doi.org/10.1145/3593434.3593478. doi:10.1145/3593434.3593478.

[13] M. T. Baldassarre, D. Gigante, M. Kalinowski, A. Ragone, Survey link, 2023. Https://forms.office.com/e/GVeeWf1Pqz.

[14] F. Zhengxin, Y. Yi, Z. Jingyu, L. Yue, M. Yuechen, L. Qinghua, X. Xiwei, W. Jeff, W. Chen, Z. Shuai, C. Shiping, Mlops spanning whole machine learning life cycle: A survey, ArXiv abs/2304.07296 (2023).

[15] W. Jin, J. Fan, D. Gromala, P. Pasquier, G. Hamarneh, Euca: the end-user-centered

explainable ai framework (2021). arXiv:2102.02437.
[16] S. Muhammad, The fairness handbook, 2022. URL: https://amsterdamintelligence.com/resources/the-fairness-handbook.
[17] ENISA, Securing machine learning algorithms, 2021. URL: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms.
[18] M. T. Baldassarre, D. Gigante, M. Kalinowski, A. Ragone, Polaris appendix, 2023. Https://figshare.com/s/1a104ceab72c73137916.
[19] ICO, Guidance on AI and data protection, 2021. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.
[20] Tensorflow, Responsible ai in your ml workflow (2021). URL: https://www.tensorflow.org/responsible_ai?hl=en.
[21] Microsoft, Threat modeling AI/ML systems and dependencies, 2021. URL: https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml.
[22] D. CSIRO, Responsible ai pattern catalogue (2022). URL: https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/.
[23] A. Wildberger, Alleviating the opacity of neural networks, in: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), volume 4, 1994, pp. 2373–2376 vol.4. doi:10.1109/ICNN.1994.374590.
[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature Machine Intelligence 2 (2020) 56–67. URL: https://doi.org/10.1038/s42256-019-0138-9. doi:10.1038/s42256-019-0138-9.
[25] M. T. Baldassarre, V. S. Barletta, G. Dimauro, D. Gigante, A. Pagano, A. Piccinno, Supporting secure agile development: The vis-prise tool, in: Proceedings of the 2022 International Conference on Advanced Visual Interfaces, AVI 2022, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3531073.3534494. doi:10.1145/3531073.3534494.