Responsibile and Reliable AI: Activities of the CINI-AIIS Lab at University of Naples Federico II

Flora Amato[1,†], Giovanni Maria De Filippis[1,†], Antonio Galli[1,†], Michela Gravina[1,†], Lidia Marassi[1,†], Stefano Marrone[1,*,†], Elio Masciari[1,†], Vincenzo Moscato[1,†], Antonio M. Rinaldi[1,†], Cristiano Russo[1,†], Carlo Sansone[1,†] and Cristian Tommasino[1,2,†]

[1]*Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, 80125, Naples, Italy*

Abstract

Over the course of the last decade, AI researchers have made groundbreaking progress in hard and longstanding problems related to machine learning, computer vision, speech recognition, and autonomous systems. Despite the success of AI, its adoption so far is mostly in low-risk applications, while the uptake in medium/high-risk applications, which might have a deeper transformative impact on our society, such as in healthcare, public administration, safety-critical industries etc., is still low compared to expectations. The reasons for such lagging are profound and range from technological limitations to difficulties associated with the conformity assessment to policies and standards. This paper introduces and discusses the perspectives and initiatives undertaken in these regards by the CINI AI-IS (the Italian National Consortium for Informatics, Artificial Intelligence and Intelligent Systems) Lab at the University of Naples Federico II.

Keywords

Artificial Intelligence, Ethics, Human-Centred, Trustworthy, Deep Learning, Machine Learning

1. Introduction

As artificial intelligence (AI) becomes increasingly integrated into critical sectors such as healthcare, finance, and transportation, the need for a more reliable and responsible deployment of AI technologies is becoming central. This widespread application underscores the necessity for regulations and certifications to manage the profound impact that AI systems are expected to, and are already having, on society and individual lives, to define the operational and developmental framework for these technologies. The current landscape of regulations governing AI is characterized by a diverse and evolving framework that varies significantly across different regions.In the European Union, the AI Act is a pioneering legislative effort that aims to set a comprehensive regulatory framework for AI, focusing on risk assessment and mitigation. It classifies AI systems according to their risk levels and imposes stricter requirements on high-risk applications, particularly in critical areas such as biometric identification and healthcare. The United States, while lacking a unified federal framework, sees regulatory initiatives that are more sector-specific and decentralized, as suggested by the AI Bill of Rights[1]. Agencies like the Federal Trade Commission (FTC) and the Food and Drug Administration (FDA) have issued guidelines that address AI's use in consumer protection and medical devices, respectively [1]. In Asia, countries like China and Singapore have also made significant strides in establishing AI guidelines, with the former working on a series of ethics guidelines and governance principles [2], focusing on controlling AI's social impacts and promoting shared norms. Singapore has been a front-runner with its Model AI Governance Framework, which provides detailed and actionable guidance to private-sector companies on responsible AI deployment [3].

Alongside governmental regulations, industry standards play a crucial role in shaping the AI regulatory landscape. Organizations such as the Institute of Electrical and Electronics Engineers (IEEE), the International Organization for Standardization (ISO) and the European Committee for Standardization/European Committee for Electrotechnical Standardization (CEN/CENELEC) have developed standards that provide frameworks for AI ethics, performance, and safety. Moreover, certifications are emerging as important tools for ensuring compliance with ethical standards and regulatory requirements, mainly with the aim of reassuring consumers, partners, and regulators of an AI system's adherence to accepted norms and practices. These processes are supported by governments across Europe, with different initiatives

---

[1]https://www.whitehouse.gov/ostp/ai-bill-of-rights/

that are actively leveraging AI to foster innovation and address societal challenges, implementing a variety of policies and funding mechanisms to support AI research, development, and integration into key sectors. Italy, in particular, is advancing its AI initiatives through the National Recovery and Resilience Plan (PNRR). This strategic plan focuses on enhancing Italy's digital infrastructure and capabilities in AI, aiming to improve public sector efficiency and drive economic growth. Investments are directed towards integrating AI in public administration, healthcare, and environmental sustainability, showcasing a robust commitment to digital transformation in line with EU priorities.

In this paper, we will thus introduce and discuss the perspectives and initiatives undertaken on responsible and reliable AI by the CINI AI-IS (the Italian National Consortium for Informatics, Artificial Intelligence and Intelligent Systems) Lab at the University of Naples Federico II, specifically focusing on the activities involving the members of the PICUS Lab[2] as part of the AI-IS Node. To this aim, Section 2 will describe the lab's activities concerning the AI certification and regulations, from both a technical and an ethical perspective, while Section 3 will introduce the FAIR project, an initiative aiming to guide frontier research for advanced AI methodologies and techniques.

2. The role of certification and regulations in AI

As highlighted in Section 1, the role of industrial standards as well as of independent certification procedures is pivotal in shaping the landscape of a resilient and reliable AI deployment. These frameworks not only ensure that AI systems operate within ethical and technical guidelines but also enhance trust and reliability in AI applications across various sectors.

2.1. The current AI standardization landscape

The current landscape of AI standardization is a dynamic and complex field characterized by efforts from various international bodies to develop and refine standards that address the rapid advancements in AI technology [4, 5]. Key organizations like IEEE, ISO, and CEN/CENELEC are at the forefront, each contributing to a global framework that aims to ensure AI systems are developed and deployed ethically and safely:

- **IEEE**: The Institute of Electrical and Electronics Engineers is a prominent entity known for setting industry standards, in various technology

fields. IEEE is also been working on initiatives around ethical considerations and safety in AI technologies. Among all, the IEEE P7000 series [6] stands out in this regard, featuring standards such as P7001 (which enhances transparency in autonomous systems), P7003 (which addresses concerns related to algorithmic bias) and P7006 (which focuses on the management of personal data by AI agents);

- **ISO**: The International Organization for Standardization, in partnership with the International Electrotechnical Commission (IEC), actively develops standards that address a wide range of issues concerning AI, such as terminology, data quality, lifecycle processes, robustness, and bias. These efforts aim to ensure the safety, reliability, and interoperability of AI systems. Notable standards include ISO/IEC 23053:2022 (which focuses on frameworks for machine learning systems), ISO/IEC TR 24027:2021 (which focuses on bias in AI systems and AI-aided decision-making) and ISO/IEC TR 24028:2020 (which details the trustworthiness of AI, covering aspects such as robustness, resilience, accuracy, and reproducibility);

- **CEN/CENELEC**: the European Committee for Standardization and the European Committee for Electrotechnical Standardization, harmonize standards across EU member states, enhancing AI technology compliance with EU norms like the AI Act. Currently, CEN/CENELEC has not published specific standards that are solely dedicated to AI. Instead, their work often integrates AI considerations into broader technological and industrial standards. They work closely with international organizations like ISO to ensure that European standards align with global efforts, particularly in areas such as data quality, security, and ethical use of technology.

2.2. Certifing AI-bases systems

Under the AI Act proposed by the European Union, national AI authorities will have significant responsibilities. Their role will include monitoring and ensuring compliance with the Act's regulations within their jurisdictions. These authorities will assess AI systems for adherence to stipulated standards, particularly for high-risk applications, ensuring that these systems do not compromise safety or public interests. Additionally, they will provide guidance to organizations on implementing AI technologies in line with the AI Act's requirements, enhancing the overall governance of AI across the EU. To support this, in the European Union several key certification authorities are responsible for ensuring compliance of industry applications with AI standards. Notably, the European

Commission itself plays a pivotal role by setting regulatory frameworks such as the AI Act. National bodies like Germany's TÜV and France's AFNOR also contribute significantly. In Italy, ACCREDIA is the central body that certifies AI according to national and EU standards, ensuring that AI systems are safe, reliable, and adhere to the required ethical guidelines. These authorities collectively uphold the integrity and trustworthiness of AI applications across Europe.

2.3. The Naples node's activities

Over the years, the Picus Lab has been active in the field of responsible and reliable AI, with applications to different domains including cybersecurity [7, 8], generative and foundation models [9, 10], law and compliance [11, 12], education [13, 14] and society [15, 16]. The lab has also been active in promoting trustworthy and human-centred AI, among which it is worth mentioning chairing workshops series co-located with important international conferences (e.g., HCAI-EP[3], HCAI4U[4]) and founding and implementing a Human-Centered AI Master's program (HCAIM[5]), a master program, co-financed by the Connecting Europe Facility of the European Union, developed by a consortium consisting of four European universities, three Centres of Excellence (CoE), and three SMEs, offering an integrated ethical, technical, and practical curriculum for understanding the construction of AI models, their realization at an industrial scale, and the evaluation of their long-term impact on society.

Beyond scholarly contributions, the lab is directly involved in a variety of actions that underscore its commitment to this vital area. These initiatives include collaborative participation with regulatory boards as well as certification agencies that aim to support reliability and responsible AI.

The lab is actively engaged with the activities of CEN/CENELEC. Specifically, one of the lab members (L.M.) has been appointed as one of the CINI AI-IS national experts for Uninfo[6], the national standardization body for Information Technologies and their applications in Italy, representing and promoting the national strategy in international standardization bodies such as CEN and ISO, as well as UNINFO in the European Telecommunications Standards Institute (ETSI). The activities are part of the working group JTC21, which focuses on the standardization of ethical and social implications of AI. Three standards are currently under development in this group: the AI Trustworthiness Framework, which will be used for third-party conformity assessments for the AI Act; Standards on Ethics, defining processes and competen-

cies for AI ethicists, being advanced towards certification in both France and Italy; and the fundamental rights impact assessment.

Concerning the activities on AI certification procedure, the lab is part of a project involving Accredia and the CINI AI-IS on the study and definition of procedures for the conformity adherence of AI-based systems to international AI technical standards. This is crucial for interoperability, safety, and ethical alignment across different industries and applications, providing a common language and expectations for developers, users, and regulators. Given the lab expertise, the activities are focused on adherence to the standard ISO/IEC TR 24027:2021, focusing on bias in AI systems and AI-aided decision-making, considering, as a case of study, the healthcare domain. To this aim, we first conducted a thorough analysis of the standard, to better frame the concept of bias, identifying its sources and potential mitigation actions from a technical perspective according to the standard itself. Subsequently, we examined the classical software lifecycle of an AI system in the medical domain to determine the optimal insertion points for compliance checks. Lastly, we proposed a procedure to check the adherence to these standards, designed to assist developers in making their products compliant, while also enabling the certification body to quantitatively verify adherence to the standards.

3. Resilient AI

The University of Naples Federico II (UNINA) leads the Spoke 3, named Resilient AI, of the Future Artificial Intelligence Research (FAIR) project, founded by Italian PNRR. Resilient AI is encompassed within the broader frameworks of responsible and reliable AI. In the context of responsible AI, which involves considering the ethical implications of AI technologies, resilient AI plays a crucial role in addressing technical risks and vulnerabilities that may compromise ethical considerations. By building AI systems that can withstand challenges and adapt to changing conditions, developers enhance the overall reliability and trustworthiness of AI technologies within an ethical framework. Similarly, within the scope of reliable AI, which focuses on building systems that consistently produce accurate, trustworthy results, resilient AI complements this objective by addressing technical challenges that may impact system reliability. These challenges include adversarial attacks, data perturbations, or system failures. By incorporating resilience into AI design, developers can enhance the robustness and dependability of AI technologies, thereby improving their overall reliability.

Spoke 3 addresses the study of AI foundational methodologies that are aimed at processing data in-the-wild,

---

[3]https://hcai-ep.sigcseire.acm.org/2024/
[4]https://sites.google.com/view/hcai4u2023
[5]https://humancentered-ai.eu/
[6]https://www.uninfo.it/

making the performance of AI resilient and robust in challenging contexts. We study how learning algorithms can cope with the problem of training with real-world data and we devise novel theories, methods, and automated instruments to address the current limitations of AI-intensive software system development, and also pay attention to the ethical and legal issues that involve AI applications in-the-wild. The research activities to be carried out include: i) the definition of appropriate data augmentation techniques, when data are incomplete or not adequately representative, while analyzing, monitoring, and improving the fairness of the machine learning algorithms; ii) the definitions of algorithms that are both resilient and robust with respect to possible external attacks (also deriving from training with "malicious" data); iii) the investigation of the implications related to the design, validation & verification, evolution and operation of the software that implements machine or deep learning algorithms, when they have to work in-the-wild; (iv) the ethical and legal issues connected with the use of real-world data.

Responsible AI endeavors to ensure that AI systems operate ethically, fairly, and transparently, with due consideration given to their societal impact [17]. In the pursuit of Responsible AI, one crucial aspect lies in the meticulous curation and semantic enrichment of datasets, a process integral to the activity of dataset recognition and semantification. Such activity delves into the creation and annotation of extensive datasets. This task is propelled by a multifaceted approach, beginning with a meticulous literature review aimed at identifying pertinent datasets across diverse domains. The distinction between general-purpose and domain-specific datasets lays the groundwork, with renowned repositories such as WordNet [18] and ImageNet [19] serving as pivotal reference points. Leveraging these gold standard datasets not only facilitates knowledge representation but also augments the semantic integration and labelling processes in several domain, such as biology, autonomous driving, and speech Processing. The guiding principles underlying this endeavor are encapsulated within the FAIR project. Drawing inspiration from these principles, the adoption of semantic artifacts and semantics-based techniques emerges as a cornerstone strategy [20, 21]. Semantification, the subsequent phase, emerges as a pivotal process aimed at imbuing data with contextual meaning through the incorporation of semantic artifacts such as ontologies and knowledge graphs. The pursuit of semantification yields manifold benefits, fostering a standardized framework for data representation, facilitating the harmonization of heterogeneous datasets, and furnishing a flexible structure for entity linkage across disparate datasets. Notably, initiatives such as the development of ImageNet++ underscore the commitment to enriching existing repositories, thereby fortifying the foundation

for subsequent AI endeavors. For specific domains, the semantic integration of standard datasets, such as for example cBioPortal[22], UniProt[23], GenBank [24] in biology, could potentially allow to discover novel insights and to make implicit knowledge explicit. Through strategic alignments with domain ontologies and meticulous mapping endeavors, the semantic labeling of datasets is poised to usher in a new era of data-centric AI. Looking ahead, the trajectory of Dataset Recognition and Semantification converges with the emergent paradigm of Responsible AI, wherein data assumes a pivotal role. By prioritizing data-centric methodologies, characterized by outlier detection, error correction, and consensus establishment, the endeavor endeavors to foster AI systems that are not only technically robust but also socially responsible and ethically sound.

The research activities of Spoke 3 also aim at addressing AI resiliency in adversarial scenarios from different points of view, towards the design of approaches and methodologies intended to i) detect and recover from attacks, ii) increase the robustness of federated learning, iii) enforce privacy, iv) enforcing fairness. Moreover, in the knowledge representation area, we will develop inference-proof countermeasures against attacks to knowledge confidentiality, based on various kinds of background knowledge and meta-knowledge.

Scenarios involving multi-task learning with missing and/or noisy labels are included with the the aim of defining effective learning procedures. In particular, in case of missing labels, the research activity will concern joint training techniques exploiting the concept of label masking or other similar approaches, while in case of noisy labels, the goal is the design of novel learning procedures optimized for soft labels, in order to take into account the uncertainty of the noisy annotations.

The need of handling missing or noisy data is also present in multimodal scenarios, where multiple data modalities should be merged to have a complete understanding of the phenomenon to be analyzed. Indeed, in several domains, such as healthcare, it is not easy to have a well-annotated dataset with paired acquisitions, consisting of samples including all the modalities. As a consequence, strategies to deal with incomplete data should be introduced, making the model robust against noisy or missing modalities. To this aim, in our research activities, we will focus on multi-input multi-output neural network, able to be flexible to the heterogeneous characteristics of the input. Moreover, in the context of multimodal learning, we will also evaluate different fusion strategies aiming to improve the integration of multiple sources.

Dealing with Resiliet AI, the Spoke 3 will provide a transformation in various aspects of our society by enabling systems and technologies to adapt, recover, and face different challenges. Indeed, Resilient AI has the

potential to drive innovation, improve resilience, and enhance societal well-being across various domains.

Acknowledgments

References

[1] A. Giovannini, A. S. Pasha, Artificial intelligence: A legal landscape, Laws of Medicine: Core Legal Aspects for the Healthcare Professional (2022) 387–404.

[2] W. Wu, T. Huang, K. Gong, Ethical principles and governance technology development of ai in china, Engineering 6 (2020) 302–309.

[3] A. A. Guenduez, T. Mettler, Strategically constructed narratives on artificial intelligence: What stories are told in governmental artificial intelligence policies?, Government Information Quarterly 40 (2023) 101719.

[4] P. Cihon, M. J. Kleinaltenkamp, J. Schuett, S. D. Baum, Ai certification: Advancing ethical practice by reducing information asymmetries, IEEE Transactions on Technology and Society 2 (2021) 200–209.

[5] M. Blösser, A. Weihrauch, A consumer perspective of ai certification–the current certification landscape, consumer approval and directions for future research, European Journal of Marketing 58 (2024) 441–470.

[6] S. Spiekermann, Ieee p7000—the first global standard process for addressing ethical concerns in system design, in: Proceedings, volume 1, MDPI, 2017, p. 159.

[7] M. Gravina, A. Galli, G. De Micco, S. Marrone, G. Fiameni, C. Sansone, Fead-d: Facial expression analysis in deepfake detection, in: International Conference on Image Analysis and Processing, Springer, 2023, pp. 283–294.

[8] L. Marassi, S. Marrone, What would happen if hackers attacked the railways? consideration of the need for ethical codes in the railway transport systems, in: Applications of Artificial Intelligence and Neural Systems to Data Science, Springer, 2023, pp. 289–296.

[9] N. Patwardhan, S. Shetye, L. Marassi, M. Zuccarini, T. Maiti, T. Singh, Designing human-centric foundation models, reconstruction 9 (2023) 10.

[10] L. Marassi, Assessing user perceptions of bias in generative ai models: Promoting social awareness for trustworthy ai, in: Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice, 2023, pp. 46–46.

[11] C. Todorova, G. Sharkov, H. Aldewereld, S. Leijnen, A. Dehghani, S. Marrone, C. Sansone, M. Lynch, J. Pugh, T. Singh, et al., The european ai tango: Balancing regulation innovation and competitiveness, in: Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice, 2023, pp. 2–8.

[12] L. Marassi, N. Patwardhan, F. Gargiulo, Can justice be a measurable value for ai? proposed evaluation of the relationship between nlp models and principles of justice (2023).

[13] F. Flammini, S. Marrone, Distance education boosting interdisciplinarity and internationalization: an experience report from "ethics, law and privacy in data and analytics" at supsi, in: Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice, 2023, pp. 54–54.

[14] B. Feeney, M. Zuccarini, T. Singh, H. Aldewereld, S. Marrone, K. Quille, Developing a human centred ai masters: The good, the bad and the ugly, in: Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 2, 2022, pp. 660–661.

[15] L. Marassi, A. E. Pascarella, G. Giacco, M. Zuccarini, S. Marrone, C. Sansone, D. Amitrano, M. Rigiroli, Artificial intelligence and voluntary carbon marketplaces: An analysis of the ethical and legal aspects, in: Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice, 2023, pp. 53–53.

[16] G. Orrù, A. Galli, V. Gattulli, M. Gravina, M. Micheletto, S. Marrone, W. Nocerino, A. Procaccino, G. Terrone, D. Curtotti, et al., Development of technologies for the detection of (cyber) bullying actions: The bullybuster project, Information 14 (2023) 430.

[17] V. Dignum, Responsible artificial intelligence: how to develop and use AI in a responsible way, volume 1, Springer, 2019.

[18] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[20] A. M. Rinaldi, C. Russo, et al., A novel framework to represent documents using a semantically-grounded graph model., in: KDIR, 2018, pp. 201–209.

[21] K. Madani, C. Russo, A. M. Rinaldi, Merging large ontologies using bigdata graphdb, in: 2019 IEEE

International Conference on Big Data (Big Data), IEEE, 2019, pp. 2383–2392.

[22] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal, Science signaling 6 (2013) pl1–pl1.

[23] U. Consortium, Uniprot: a hub for protein information, Nucleic acids research 43 (2015) D204–D212.

[24] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, Genbank, Nucleic acids research 41 (2012) D36–D42.