Symbiotic AI: What is the Role of Trustworthiness?

Miriana Calvano[1], Antonio Curci[1,2], Rosa Lanzilotti[1] and Antonio Piccinno[1]

[1]University of Bari "Aldo Moro", Via Edoardo Orabona 4, 70125, Bari, Italy
[2]University of Pisa, Largo B. Pontecorvo 3, 56127, Pisa, Italy

Abstract

The design, development, and use of Artificial Intelligence (AI) is crucial in modern society. The traditional design of AI systems focuses on models with very high performances without highlighting how relevant the role of humans is in this context. To create AI systems that suit end users' needs and preferences, it is important to involve them in each phase of the system lifetime cycle. AI systems must present interfaces and interaction paradigms that enhance users' cognitive models, ensuring usability and a positive User Experience (UX). In this new scenario, Human-Computer Interaction (HCI) and AI contaminate each other leading to reach the human-AI symbiosis. Researchers should shift the focus toward Symbiotic AI (SAI) systems, which aims to enhance humans' abilities without replacing them. This manuscript presents preliminary considerations for the creation of a framework to design high-quality SAI systems and metrics that can be employed to appropriately evaluate them. Being a novel field, it focuses on the current investigation regarding the definition of the properties of SAI systems, stressing the importance of Trustworthiness, and whether new design principles for SAI systems can be extracted from the AI act.

Keywords

Symbiotic AI, Trustworthiness, Design and Evaluation, Human-Centered Approach, AI Act (AIA)

1. Introduction

The fast and broad spread of artificial intelligence (AI) over the past few years has allowed individuals to use new services, products, and systems to perform various tasks and activities. AI has been introduced in various fields, such as medicine, law, and education, raising several concerns because the results of the systems can influence humans to make decisions that are often irreversible and can impact other individuals. Consequently, legal bodies and governments are working to regulate AI to preserve humans with new laws, such as the Artificial Intelligence Act (AIA), which undertakes a risk-based approach regarding the design, development, and deployment of AI for EU citizens, identifying its best and forbidden practices while delineating guiding principles [1]. This implies that the future direction of AI is undergoing substantial changes that should be addressed with a multidisciplinary approach [2].

The main issue with AI systems is that the traditional approach to their development heavily focuses on achieving high-performing models and obtaining excellent metrics (e.g., accuracy, precision, recall). Such models are also called *black boxes*: users cannot analyze and comprehend the processes that lead to the outputs of such systems, causing low transparency [3]. This can be addressed by adopting a human-centered approach when designing and developing AI systems to foster a symbiotic relationship with humans and let technology support humans' daily activities without replacing them, adapting to their mental and physical models [4]. Human-Centred Design (HCD), which belongs to the Human-Computer Interaction (HCI) discipline, stresses that end-users must always be involved in the creation of any kind of product, in order to create clear, appropriate and effective interfaces that allow end-users to interact correctly with the software they are using [5, 6, 7, 4]. On the other hand, software engineering (SE) is another pillar in the development of quality software systems, as it is the discipline that studies how software should be developed, maintained and used through specific standards and processes [8]. It is, therefore, crucial to integrate practices and principles from the two disciplines to support designers and developers in creating artificial intelligence systems that enable a symbiotic relationship with their end users.

This research is part of the *Future Artificial Intelligence Research (FAIR)* project, which aims to bring innovation to the European Union in the context of AI. FAIR follows a holistic and multidisciplinary approach to rethink the foundations of AI and investigate its social impact. Its goal is to build systems capable of interacting and collaborating with humans and foster trustworthiness. Specifically, the research presented in this article is performed within the Spoke 6, named Symbiotic AI (SAI), which investigates the scientific, social, economic, legal and ethical challenges related to the growing symbiosis between humans and AI. SAI refers to a collaborative re-

Ital-IA 2024, 29-30th May 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
miriana.calvano@uniba.it (M. Calvano); antonio.curci@uniba.it (A. Curci); rosa.lanzilotti@uniba.it (R. Lanzilotti); antonio.piccinno@uniba.it (A. Piccinno)
0000-0002-9507-9940 (M. Calvano); 0000-0001-6863-872X (A. Curci); 0000-0002-2039-8162 (R. Lanzilotti); 0000-0003-1561-7073 (A. Piccinno)

lationship between humans and AI systems in which "the human understands and intuitively reacts to the machine, and the machine understands and intuitively reacts to the human" [9].
To reach the human-AI symbiosis, users should trust the system's decisions and properly comprehend them, making *Trustworthiness* one of the main properties to consider when dealing with such systems. However, due to the novelty of the field, limited work is available in the literature. Our research aims to propose a comprehensive framework and evaluation metrics to support designers, developers, and AI specialists in creating and evaluating Symbiotic AI (SAI) systems that inspire trust, ensure fairness, and are responsible and compliant with the various domains in which they operate [10].

This manuscript is structured as follows: Section 3 describes the approach that will be undertaken to design and evaluate SAI systems; Section 2 presents how trustworthiness can be defined in the SAI field, exploring the perspectives of the European Commission and academia; Section 4 concludes and explores the future work of the project.

2. Trustworthiness for SAI Systems

For people and society, *trustworthiness* is undoubtedly one of the prerequisites that AI systems should have to be used without hesitations [11]. It, therefore, becomes the starting point of our research because of its breadth and multifaceted nature. In this section, the concept of trustworthiness is explored by analysing the perspectives of European policymakers and academics to determine how to consider it in the context of SAI.

2.1. The European Commission Perspective

This section focuses on two documents drafted by the European Commission: the Ethics Guidelines for Trustworthy AI and the AIA. The goal is to delineate a clear image of the standpoints of policymakers to create AI products that fully comply with laws, regulations, and norms and track the efforts of the EU concerning human rights, ethics, and philosophical issues.

2.1.1. Ethics Guidelines for Trustworthy AI)

The role of the AI HLEG is to define the approach of the European Commission with respect to AI by indicating the key principles and policies. In 2019, they drafted the "Ethics Guidelines for Trustworthy AI" report, which identifies seven requirements of *Trustworthiness*, identified as the umbrella property to ensure a human-centric approach to AI [11, 12], illustrated in Figure 1. Such requirements are briefly described in the following:

- **Human agency and Oversight**: incorporating mechanisms for human intervention in critical decision-making processes ensures human control and supervision over AI systems to prevent unintended consequences.
- **Technical Robustness and Safety**: developing AI systems necessitates a risk-preventive approach that ensures reliable behavior, minimizing and preventing unintentional and unexpected harm.
- **Privacy and Data Governance**: ensuring privacy protection requires robust data governance, encompassing both the quality and integrity of the data used in processing to guarantee privacy.
- **Transparency**: encompassing the transparency of elements requires to comprehend the reason that lies behind the decision taken by the system.
- **Diversity, Non-Discrimination and Fairness**: involving all stakeholders throughout the entire system lifecycle ensures equal access through inclusive design processes and equitable treatment.
- **Societal and Environmental Well-being**: maximizing sustainability, social impact, and ecological responsibility of AI systems to positively contribute to society while minimizing negative consequences.
- **Accountability**: creating mechanisms to ensure accountability of AI systems, both before and after their development, deployment and use guarantees fairness [11].

2.1.2. The Artificial Intelligence Act (AIA)

Starting from the requirements of Trustworthy AI, listed in Section 2.1.1, in 2021, the EU has defined the AIA to regulate the adoption of harmonised and standardized rules for AI systems. Specifically, it merges trustworthiness with the risk-based approach to determine the acceptability of the types of systems through norms and regulations [12]. The risk-based approach outlines four categories of AI systems in relation to the risks they might cause:

1. *Unacceptable Risk*: it encompasses systems that might include prohibited AI practices that must be banned to guarantee a well-functioning society, such as those that might threaten minorities or those used by public authorities.
2. *High Risk*: it regards systems used in fields such as education and vocational training, access to private and public services, law enforcement, etc.

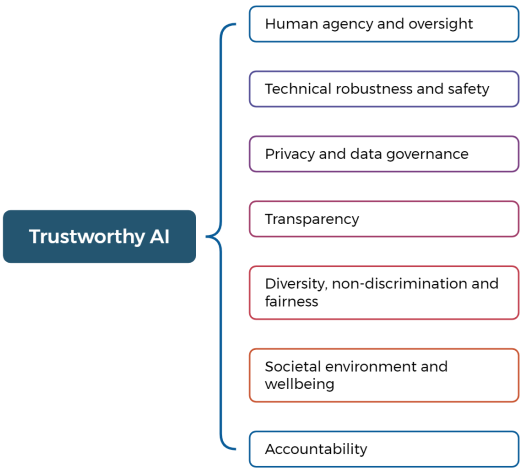

Figure 1: The seven key requirements of Trustworthy AI: all are of equal importance and support each other [11]

3. *Limited Risk*: it encompasses AI systems that must comply with specific transparency obligations because they interact with humans (e.g., biometric recognition systems, and emotion recognition systems).
4. *Low or Minimal Risk*: it refers to systems that feature AI but do not require specific conformity checks [1].

2.2. The Academic Perspective

Ben Shneiderman, one of the pioneers of HCI, proposes trustworthiness as one of three principles, along with safety and reliability, of human-centered AI (HCAI) systems, which guarantee an appropriate balance of automation and human control. Specifically:

- *Trustworthiness* concerns the property that makes systems deserving of being trusted by humans.
- *Reliability* comes from the application of technical practices of software engineering that build systems that produce appropriate and/or expected responses.
- *Safety* is a strategy to guide the refinement of the model performance to prevent potential failure and improper use [13].

The three above mentioned properties are the most recurrent in the literature since they are the main areas of research and can encompass the other properties; nevertheless, the state of the art concerning the human-AI interaction, considers other 22 properties that can influence the design and development of any kind of system (e.g., usable, observable, explainable, resilient, agile, etc.) [13].

The investigation of our research work consists in understanding what principles are applicable to SAI and identifying the potential new properties.

2.3. The Impact of Trustworthiness in SAI

Our objective is to define a framework that encompasses both standpoints; in this regard, the authors are performing a Systematic Literature Review (SLR), following the Kitchenham protocol, to identify the guidelines and principles that can be drawn from the AIA that could be applied to the lifecycle of SAI systems [14]. This SLR has the objective to determine how the research community is investigating and employing the AIA with respect to the design and development AI. From the preliminary results, it emerged that trustworthiness is intrinsic in SAI because humans must fully trust systems in order to symbiotically with them.

Belonging to the domain of AI built following a human-centered approach, SAI can include *Trustworthiness*, *Safety*, and *Reliability* as principles; however, the establishment of a symbiotic relationship might require their refinement or to the definition of new ones. The ongoing SLR will also serve to establish the new principles and identify new guidelines suitable for the field of SAI.

3. Conceptual Framework for SAI Systems

The starting point is understanding the gaps in the traditional approach to the development of AI systems to determine the changes to propose and the integration of new processes into the software lifecycle. This conceptual framework aims to support designers and developers in creating and evaluating SAI systems. The objective is to provide a standardized methodology to those who create AI-powered services that reduce the gap between technology and humans and decrease cognitive demand when interpreting and understanding the outputs that systems produce.

The objective of this work lies in defining a framework that considers and merges the two perspectives (i.e. Ethics Guidelines for Trustworthy AI and AIA), while identifying principles, guidelines, and techniques that belong to different disciplines by finding the appropriate links. Figure 2 presents an initial version of the conceptual framework that consists in two layers, *Design* and *Assessment*, explained below.

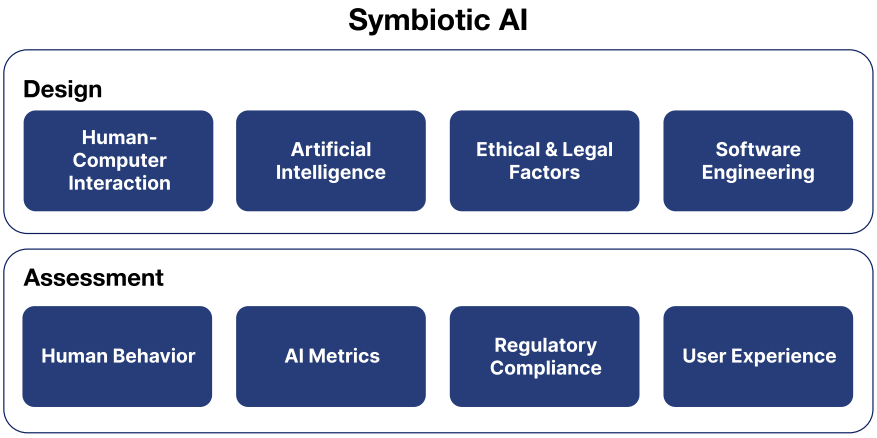

Figure 2: Conceptual Comprehensive Framework for the design and the evaluation of Symbiotic AI

3.1. Design

This layer embraces four main research areas that contribute equally: Human-Computer Interaction (HCI), Law & Ethics, Software Engineering (SE), and AI. The following sections describe each component of the framework, illustrating its role in the SAI scenario.

Human-Computer Interaction (HCI) HCI is one of the pivotal components of this framework because the symbiotic relationship can be achieved if such systems allow users to reach their goals with effectiveness, efficiency, and satisfaction, thus, by being usable and providing a positive user experience. Other key elements that HCI is responsible for are feedback and affordance, enabling humans to understand how the system should be used, making them feel at ease with proper communication [6]. Involving humans iteratively during each phase of the system's lifecycle implies performing interviews, questionnaires, field studies, and focus groups to perform quantitative and qualitative evaluations of the systems and to obtain rich insights concerning the users' needs, preferences and cognitive models [6, 7].

Ethical & Legal Factors This dimension considers the regulatory, philosophical, and ethical standpoint since designers and developers must create products that preserve users' social, working, and personal well-being. One of the main issues concerning AI, which becomes particularly valid for the branch of SAI, consists of avoiding biases and ensuring fairness. This element must be always considered because the root of biases is found in how data is treated by AI models, for example, in the learning phase. This determines the unfair behavior of systems that can influence humans' decisions when employing AI as an instrument. The legal standpoint must be considered for designing and developing AI systems to create products that comply with regulations and can be released to the public. Currently, the main elements to consider are the AIA and the General Data Protection Regulation (GDPR); the first regulates the design, development, and use of AI systems in the EU, while the GDPR is a law that defines how data is handled, stored, and processed [15].

These regulations define the ethical principles that any kind of system should possess to be available to society.

Artificial Intelligence (AI) This dimension refers to AI from a technical and algorithmic standpoint because the framework aims to suggest the appropriate techniques and practices to adopt depending on the requirements of the systems to create. AI models, along with high computational power, can be employed in multiple domains, such as business, finance, healthcare, agriculture, smart cities, and cybersecurity; however, they cannot be used as a one-size-fits-all solution because, depending on the activities, different tasks are needed - e.g., classification, prediction, description -, raising the need for context-specific models, parameters, and variables [16]. The effectiveness of SAI systems is not guaranteed by simply obtaining high-performing models but rather by systems that properly integrate *Transparency*, *Explainability*, and *Interpretability*. This provides users with the right instruments to comprehend the processes behind outputs, influencing their decisions, and what data is responsible for the system's responses.

Software Engineering (SE) This framework aims to guide design and developers in creating SAI systems, ensuring that they operate by following a human-centered

approach while complying with legal requirements and implementing high-performing AI systems. Thus, the objective is to integrate the Agile principles and the processes of the Agile Development Lifecycle with those belonging to the SAI design, creating a mapping that does not exclude any discipline [17].

3.2. Assessment

In this new scenario, where a strict correlation and contamination exists between human and AI performance, it becomes essential to define novel metrics to assess the human-AI symbiotic relationship.

Traditionally, human beings and AI have been viewed as distinct and unrelated entities, causing UX and AI metrics to be defined independently to evaluate both human behavior and system performance. Considering them in unison, it is possible to draft a preliminary set of metrics that can be employed to assess the symbiosis. By integrating both the dataset and user information and considering the user's characteristics from the training phase of the AI model, it is possible to foster symbiosis, making the system's behaviour as much as possible adaptable to the user's needs.

Since *Trustworthiness* allows users to trust systems that operate safely and exhibit reliable behavior, it is contemplated as one of the starting points of this research work [4]. Assessing this aspect is difficult since it varies across many application contexts [4]; therefore, it is necessary to understand whether its evaluation should consider it as a stand-alone property or as an ensemble of other dimensions, such as safety, fairness, robustness, etc[1].

Two potential metrics are proposed to assess how Trustworthy an AI system is: *Preventing Undesired System Behaviors*, which refers to how effectively the system avoids actions that could potentially harm the user or deviate from expected behavior; *Correctness of Decisions*, which measures the extent to which system's decisions align with user expectations and desired outcomes.

4. Conclusions

This paper presents preliminary considerations concerning the novel field of Symbiotic AI with respect *Trustworthiness*. It presents the main challenges of identifying the principles of this field while stressing the need for a human-centered approach when dealing with AI systems of any kind. This research work is the starting ground for the definition of a comprehensive framework, presented in Section 3, that encompasses multiple disciplines and aims to guide designers and developers in creating SAI systems. This framework is still in its early stages and at a conceptual state. Delineating a standardized approach to assess the behavior and performance of such systems is crucial to ensure the proper deployment of AI, which is part of the daily lives of countless individuals. As *Trustworthiness* plays a pivotal role in an effective human-AI interaction, the future of this research will focus on determining its complementary principles and its impact on symbiosis by carrying out verticalized case studies and performing in-depth investigations in the literature.

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Acknowledgments

The research of Miriana Calvano and Antonio Curci is supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies, and research D.D. MUR n. 341 del 15.03.2022 – Next Generation EU (PE0000013 – "Future Artificial Intelligence Research – FAIR" - CUP: H97G22000210007).

References

[1] T. E. Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on asrtificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2024. URL: http://thomas.loc.gov/cgi-bin/query/z?c102:H.CON.RES.1.IH.

[2] C. Sanderson, D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson, D. Hansen, Ai ethics principles in practice: Perspectives of designers and developers, IEEE Transactions on Technology and Society 4 (2023) 171–187. URL: http://dx.doi.org/10.1109/TTS.2023.3257303. doi:10.1109/tts.2023.3257303.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51 (2019) 1–42. URL: https://dl.acm.org/doi/10.1145/3236009. doi:10.1145/3236009.

[4] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, N. Diakopoulos, Designing the User Interface: Strategies for Effective Human-Computer Interaction, 6 ed., Pearson Education, 2016. URL: https://books.google.it/books?id=PpItDAAAQBAJ.

[5] I. O. for Standardization, Iso 9241:210 - ergonomics of human-system interaction, 2019. URL: https://www.iso.org/standard/77520.html.

[6] H. Sharp, J. Preece, Y. Rogers, Interaction Design: beyond human-computer interaction, 5 ed., John Wiley & Sons, Inc., 2019.

[7] I. O. for Standardization, Iso 9241:210 - ergonomics of human-system interaction: Human-centred de-

sign for interactive systems, 2019. URL: https://www.iso.org/standard/77520.html.

[8] P. Bourque, R. E. Fairley (Eds.), SWEBOK: guide to the software engineering body of knowledge, version 3.0 ed., IEEE Computer Society, Los Alamitos, CA, 2014.

[9] S. S. Grigsby, Artificial intelligence for advanced human-machine symbiosis, in: D. D. Schmorrow, C. M. Fidopiastis (Eds.), Augmented Cognition: Intelligent Technologies, Springer International Publishing, Cham, 2018, pp. 255–266.

[10] M. Vahabava, The risks associated with generative AI apps in the European Artificial Intelligence Act (AIA), in: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), CEUR Workshop Proceedings, Munich, Germany, 2023, pp. 1–12.

[11] E. Commission, European commission - ethics guidelines for trustworthy ai, 2021. URL: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[12] J. Laux, S. Wachter, B. Mittelstadt, Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, Regulation & Governance 18 (2024) 3–32. URL: https://onlinelibrary.wiley.com/doi/10.1111/rego.12512. doi:10.1111/rego.12512.

[13] B. Shneiderman, Human-Centered AI, 1 ed., Oxford University PressOxford, 2022. URL: https://academic.oup.com/book/41126. doi:10.1093/oso/9780192845290.001.0001.

[14] B. Kitchenham, Procedures for Performing Systematic Reviews, Technical Report, Keele University, 2004.

[15] Gazzetta Ufficiale dell'Unione Europea, General Data Protection Regulation (GDPR): Regulation (EU) 2016/679, 2018.

[16] I. Sarker, Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems, 2022. doi:10.20944/preprints202202.0001.v1.

[17] D. Salah, R. F. Paige, P. Cairns, A systematic literature review for agile development processes and user centred design integration, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM, London England United Kingdom, 2014, pp. 1–10. URL: https://dl.acm.org/doi/10.1145/2601248.2601276. doi:10.1145/2601248.2601276.