Designing Symbiotic AI through a Multidisciplinary Framework

Antonio Curci[1,2]

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

Abstract

Artificial Intelligence (AI) is changing how we carry out our daily activities, impacting modern society on multiple levels. Although its use brings numerous benefits and advantages, some risks must be considered when interacting with AI-based systems, especially in high-risk situations. Collaborating with humans instead of replacing them is the goal of Symbiotic AI, which derives from the field of Human-Centred AI and focuses on a mutual exchange between humans and machines without undermining their judgement and expertise. A three-year Ph.D. project, which revolves around SAI, is presented for the creation of a multidisciplinary framework that can guide its design and development. The project is in its second phase, and the preliminary results are presented, along with considerations concerning the future of the research.

Keywords

Artificial Intelligence, Human-Centered Design, Human-Computer Interaction

1. Introduction

In a world that becomes more and more digitalized, Artificial Intelligence (AI) is at the forefront of innovation in many domains. Society is benefiting from this revolution thanks to the high computational power of AI-based systems that can elaborate, classify, and generate huge amounts of data. At the same time, several limitations and risks come from the use of such systems, especially when they are used to make decisions that impact the lives of other individuals [1]. This highlights the importance of undertaking a human-centered approach when designing and developing AI-based systems, enabling the establishment of continuous collaboration with humans and a mutual exchange for improvement of the two parties, leading to a symbiotic relationship. These characteristics are the pillars of a new branch of AI called Symbiotic AI (SAI), a subset of Human-Centred AI, that aims at augmenting humans instead of replacing them [2, 3].

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 fostering trustworthiness. Specifically, the research presented in this article is performed within Spoke 6, named, in fact, Symbiotic AI (SAI). FAIR sets the main scope of the research and the main topics that it should revolve around, the Ph.D. project focuses on the design of SAI. Thus, the contribution of the candidate's research consists of defining methodologies and techniques that allow reach this goal. More specifically, the objective of the 3-year Ph.D. project is to create a multidisciplinary framework to guide practitioners—designers and developers—in creating high-quality SAI systems. Currently, the Ph.D. project is in its second year and the framework is currently being created, composed of principles, properties, and guidelines for SAI; these elements are being validated and refined through case studies. Starting from the main research questions (see section 2), the tasks and objectives for each year are presented, along with the

CHItaly 2025: Technologies and Methodologies of Human-Computer Interaction in the Third Millenium, Doctoral Consortium, 6-10 October 2025, Salerno, Italy
antonio.curci@uniba.it (A. Curci)
https://ivu.di.uniba.it/people/curci (A. Curci)
0000-0001-6863-872X (A. Curci)

preliminary results.

This manuscript is structured in the following way: section 2 illustrates the objectives, the research questions, and the phases of this Ph.D. project; section 3 explores the state of the art concerning SAI and defines the gap that this research aims to fill; section 4 reports the current results.

2. Motivations and Objectives

Traditionally, AI systems were created by merely considering their performance: developers used to work towards the achievement of models with high accuracy, focusing on computation and architectural optimization. These aspects remain crucial, but they must be accompanied by the design of interaction mechanisms and paradigms that meet users' requirements, mental models, and preferences. This translates into the need for the application of Human-Centred Design (HCD) in its entirety [4], which stresses the importance of including end users in the creation process from the beginning to the deployment of a product [5].

As SAI is a new field of research, the literature is currently lacking a standardized approach that can be systematically followed by designers and developers when creating such systems. Thus, the outcome of this Ph.D. project is to refine the processes of the HCD approach in order to make them more appropriate for this context through the framework in question. This implies collaboration among different disciplines, even outside computer science, since the use of AI has become particularly broad and cross-domain, to focus on the mere performance of models. The contribution of this project is to define and delineate the guidelines and design patterns that support designers and developers in achieving proper architectural solutions for SAI. Based on these objectives, the research questions that guide the research are the following:

RQ1) How can the processes of the HCD be refined to foster the creation of SAI systems?

RQ2) What are the design patterns to integrate into a framework for SAI systems?

The project is conceptually divided into three phases—*Phase 1*, *Phase 2*, and *Phase 3*—which are detailed below. It is currently in the middle of its second phase, validating the mostly-theoretical results from the first year.

Phase 1 This phase consisted of studying the state of the art concerning AI systems, focusing on those that highlight the collaboration with humans. Two Systematic Literature Reviews (SLR) were carried out, which aimed at outlining the current techniques, practices, and methodologies used in this field. The first SLR had the goal of defining the principles of human-AI symbiosis with respect to the novel legal scenario of the AI Act. The output is a principle-based framework that merges HCD and the regulatory approach by defining dimensions and properties of SAI systems that comply with this EU law. The second SLR, which is more general, consisted of defining the factors that can influence the establishment of a symbiotic relationship between humans and AI, abstracting from the AI Act. This phase permitted to lay the groundwork for *Phase 2*, delineating the disciplines that contribute to building SAI systems. In addition, initial experiments were conducted to understand the behavior of SAI systems in real-world scenarios, which are being continued in the current phase. The results are better illustrated in section 4

Phase 2 It encompasses the finalization of the results of *Phase 1* and the creation of the knowledge base of the framework, listing and specifying the first set of guidelines and design patterns. The principles and properties of the AI-Act based framework are being expanded and specialized with the aid of case studies and experiments to determine their validity and refine them. It involves their application by re-engineering existing AI-based systems and/or creating new ones. Three case studies are being conducted, both in academia and in a company, investigating two domains—medicine and software engineering. This phase has not concluded yet, but the preliminary results are discussed in section 4.

Phase 3 It will focus on conducting experiments and refining the results of *Phase 2* in order to finalize the framework iteratively. The framework evaluation has two main aspects. The first focuses on assessing its applicability for designers and developers. The second involves evaluating the SAI systems developed using the framework. This second aspect will be addressed by a parallel project on the same topic, which is dedicated to defining metrics and techniques for achieving symbiosis [6].

3. Background literature and State of the Art

The collaboration and intersection between AI and Human-Computer Interaction (HCI) has consistently strengthened in the last few years as the application of HCI methodologies and practices reinforces the higher objective of creating products that support humans. Any system should allow users to grasp how to use it properly while providing them with appropriate information about the consequences of their actions [7]. The ultimate goal is the achievement of positive interactions that allow users to successfully carry out their activities, focusing on augmentation rather than automation. It is important to make sure that the system exhibits its intelligent behavior in a suitable way for its user, ensuring that its usability and utility are not compromised, fostering adaptability [8]. In SAI, End-User Development (EUD) represents another key point because ensuring that humans can build and modify the systems that they use allows them to deploy software that aligns with their expectations and desires while making them feel in control [9]. When dealing with AI-based systems, applying the guidelines and principles belonging to HCI can bring additional and domain-specific challenges to face. Unfortunately, AI is commonly mystified due to the inability of end users to understand its complexity and mathematical foundations. This decreases trust, and individuals are more unwilling to rely on it; this makes the creation of effective communication mechanisms a crucial part of designing AI, which has direct implications on the ethical and legal components [10].

The risk of facing negative irreversible consequences from wrong decisions made by AI can be minimized by ensuring that proper communication mechanisms are integrated; humans must be able to fully understand and comprehend the outputs of such systems in order to reach proper outcomes [11]. The communication in question can be achieved with transparency and explainability, which focus on providing insights into the structure of the AI model and the processes that led to specific outputs. The main challenge resides in the explanation of responses generated by *black-box* models, which are too deep and complex to be transparent or explainable [12]. This issue is still being explored in research since these two characteristics of AI-based systems highly influence their *trustworthiness* [3], a property often under debate. It constitutes an important factor in the interaction process because it must be properly balanced: over-trust can be dangerous to human agency and decision-making abilities, while under-trust can undermine the purpose which an AI system was designed for [13, 14]. Driven by these motivations, over the past few years, academia and governmental bodies have joined forces to create AI that we can trust, laying the foundation for the creation of long-term sustainable AI-based products that work in symbiosis with their end users and have a positive impact on society. In this regard, the AI Act, the first European Union legal framework for these systems, is at the core of this research. It undertakes a human-centric and risk-based approach, defining new requirements for designers and developers, as it stresses the importance of transparency in order to protect citizens' safety, security, and overall well-being [15].

4. Results

The results of the first phase of the research consist of a set of properties and dimensions that come from studying the state of the art concerning the AI Act through an SLR.

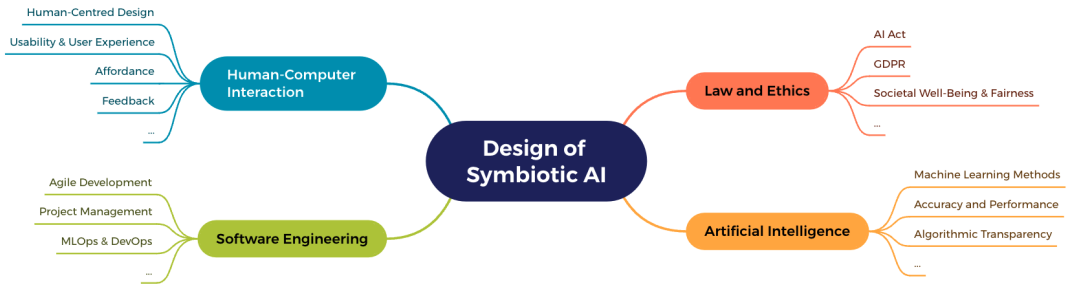

Figure 1: Conceptual multidisciplinary framework to design SAI systems

4.1. Theoretical Groundwork

The output of the SLR, presented in [16], established the need for transparency obligations, appropriate interaction mechanisms, and ensuring human control and oversight. AI must users distinguish situations where we can trust the technology from those where our judgment is necessary. This allows our common sense and experience to complement the mathematical complexity of AI, and vice versa, because there are situations and contexts in which humans need or wish for fully automated systems, in which their control is not necessary. At the same time, they must always be allowed to be in control in case they are willing to modify the AI's behavior [3]. The conceptual version of the framework, shown in Figure 1, illustrates four disciplines that are involved in the creation of SAI: *Human-Computer Interaction* stands at the intersection among the technical aspects of computer science and psychology to create intuitive, usable, and accessible AI-based systems. The latter must follow the standards and methods researched in the field of *Artificial Intelligence*, which focuses more on the mathematical and computational sides of the models. *Software Engineering* encompasses the practices and the methodologies that designers and developers are required to adhere to while complying with the legal obligations set by *Law and Ethics*.

The case studies are being carried out to investigate how the principles and guidelines can actually apply in real-world scenarios and obtain more insights concerning the best practices to reach human-AI symbiosis.

4.2. SAI in Real-World Contexts

The theoretical foundations for SAI investigated in the first phase are being employed and refined through the creation and evaluation of AI systems for different purposes. The three main case studies are reported below.

LLMs for Usability Testing The first case study concerned the employment of Large Language Models (LLMs) in HCD, specifically, in the definition of usability studies. Three general-purpose LLM-based platforms were used—ChatGPT, Mistral, and Gemini— to create the protocol and the tasks for a usability study to conduct on an AI-based system for the rhinocytology [17]. From this experiment, it emerged that LLMs can support practitioners in defining the tasks but only to a very limited extent, for example, in the brainstorming phase [17]. This case study is being carried out again on other models, with different characteristics and with different prompts that could fill this gap. This case study is a collaborative effort, in which each researcher contributes in every step and phase, but the core design choices are performed based on the output of the candidate's research of the first year.

Multimodal Models for Brain Tumor Detection The second case study consists of the design and development of a new AI model, presented in [18], in which the candidate is the principal investigator concerning design choices, data retrieval, and development. It is a multimodal neural network that classifies grayscale-2D brain tumor MRI scans based on a binary class: *ill* or *healthy* [18]. The model that was built exhibited promising results (with 90% accuracy), and is not being integrated in a fully-

functioning system that can actually establish a symbiotic relationship with humans through an interaction paradigm that was defined based on the guidelines previously mentioned. Thus, the experiment had the objective of validating the principles and deriving guidelines to create a SAI system. This case study strongly focuses on achieving the proper level of automation with respect to providing appropriate explanations. The future work will involve the integration of Reinforcement Learning (RL), through Interactive Machine Learning (IML) mechanisms, to allow physicians to correct the behavior of the model [19, 20].

Natural Language Explanations for Alzheimer's Disease The third case study is still on-going and being carried out in a company, *Lutech S.p.A.*, in which my role is to design and develop an LLM-based component of an AI system for Alzheimer's Disease detection to complement visual explanations of the Convolutional Neural Netowrk (CNN) created with Gradient-weighted Class Activation Mapping (GradCAM) [21]. This description acts as a caption of the diagnosis, providing further details on the reasoning behind the AI model, reinforcing the pillar of symbiosis, for which humans must be supplied with all the necessary instruments for understanding AI. Although the overall objective was set collaboratively with those involved in the project, the candidate was in charge of defining modalities, materials, and methods of the experiment, which was carried out with three on-premise multimodal LLMs, whose outputs are being validated with real physicians in order to determine which model, prompt, and temperature suits best this task.

5. Conclusions

The objective of this Ph.D. Project is to create a multidisciplinary framework, based on the core processes and practices of the HCD, to support designers and developers in the creation of SAI. Currently, the research is in the middle of its second phase and will be followed by the last and third one in a few months. The latter will involve the final validation and refinement of the results obtained with the theoretical and practical investigations presented in this manuscript.

Acknowledgments

The research of 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).

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] W. Xiong, H. Fan, L. Ma, C. Wang, Challenges of human—machine collaboration in risky decision-making, Frontiers of Engineering Management 9 (2022) 89–103. URL: https://link.springer.com/10.1007/s42524-021-0182-0. doi:10.1007/s42524-021-0182-0.

[2] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, M. F. Costabile, From human-centered to symbiotic artificial intelligence: a focus on medical applications, Multimedia Tools and Applications (2024). URL: https://link.springer.com/10.1007/s11042-024-20414-5. doi:10.1007/s11042-024-20414-5.

[3] B. Shneiderman, Human-Centered AI, 1 ed., Oxford University PressOxford, Great Clarendon Street, Oxford, OX2 6DP, United Kingdom, 2022. URL: https://academic.oup.com/book/41126. doi:10.1093/oso/9780192845290.001.0001.
[4] H. Sharp, J. Preece, Y. Rogers, Interaction Design: beyond human-computer interaction, 5 ed., John Wiley & Sons, Inc., 2019.
[5] I. O. for Standardization, Iso 9241:210 - ergonomics of human-system interaction: Human-centred design for interactive systems, 2019. URL: https://www.iso.org/standard/77520.html.
[6] M. Calvano, Techniques and Methods to Evaluate Human-Centered Symbiotic AI Systems, in: Companion Proceedings of the 30th International Conference on Intelligent User Interfaces, ACM, Cagliari Italy, 2025, pp. 232–234. URL: https://dl.acm.org/doi/10.1145/3708557.3716153. doi:10.1145/3708557.3716153.
[7] D. Norman, The Design of Everyday Things, revised and extended edition ed., Basic Books, 2013.
[8] A. Bunt, C. Conati, J. McGrenere, Mixed-initiative interface personalization as a case study in usable ai, AI Mag. 30 (2009) 58–64. URL: https://doi.org/10.1609/aimag.v30i4.2264. doi:10.1609/aimag.v30i4.2264.
[9] B. R. Barricelli, F. Cassano, D. Fogli, A. Piccinno, End-user development, end-user programming and end-user software engineering: A systematic mapping study, Journal of Systems and Software 149 (2019) 101–137. URL: https://linkinghub.elsevier.com/retrieve/pii/S0164121218302577. doi:10.1016/j.jss.2018.11.041.
[10] R. Parasuraman, V. Riley, Humans and Automation: Use, Misuse, Disuse, Abuse, Human Factors: The Journal of the Human Factors and Ergonomics Society 39 (1997) 230–253. doi:10.1518/001872097778543886.
[11] R. Guidotti, A. Monreale, D. Pedreschi, F. Giannotti, Principles of Explainable Artificial Intelligence, Springer International Publishing, 2021, pp. 9–31. URL: https://link.springer.com/10.1007/978-3-030-76409-8_2. doi:10.1007/978-3-030-76409-8_2.
[12] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, K. Sycara, Transparency and Explanation in Deep Reinforcement Learning Neural Networks, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, New Orleans LA USA, 2018, pp. 144–150.
[13] J. Robertson, C. Ferreira, E. Botha, K. Oosthuizen, Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction, Business Horizons 67 (2024) 499–510. URL: https://linkinghub.elsevier.com/retrieve/pii/S0007681324000533. doi:10.1016/j.bushor.2024.04.008.
[14] B. Botero Arcila, AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?, Computer Law & Security Review 54 (2024) 106012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0267364924000797. doi:10.1016/j.clsr.2024.106012.
[15] E. P. . C. of the European Union, Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828, 2024.
[16] M. Calvano, A. Curci, G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Building Symbiotic AI: Reviewing the AI Act for a Human-Centred, Principle-Based Framework, 2025. URL: http://arxiv.org/abs/2501.08046. doi:10.48550/arXiv.2501.08046, arXiv:2501.08046 [cs].
[17] M. Calvano, A. Curci, R. Lanzilotti, A. Piccinno, A. Ragone, Leveraging Large Language Models for Usability Testing: a Preliminary Study, in: Companion Proceedings of the 30th International Conference on Intelligent User Interfaces, ACM, Cagliari Italy, 2025, pp. 78–81. URL: https://dl.acm.org/doi/10.1145/3708557.3716341. doi:10.1145/3708557.3716341.
[18] A. Curci, A. Esposito, Detecting Brain Tumors Through Multimodal Neural Networks, in: 13th International Conference on Pattern Recognition Applications and Methods, SCITEPRESS Science and Technology Publications, Lda., Rome, Italy, 2024, pp. 995–1000. doi:10.5220/0012608600003654.
[19] R. S. Sutton, A. G. Barto, Reinforcement learning: an introduction, Adaptive computation and machine learning series, second edition ed., The MIT Press, Cambridge, Massachusetts, 2018.

[20] N. A. Wondimu, C. Buche, U. Visser, Interactive Machine Learning: A State of the Art Review, 2022. URL: http://arxiv.org/abs/2207.06196. doi:10.48550/arXiv.2207.06196, arXiv:2207.06196 [cs].

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision 128 (2020) 336–359. URL: http://link.springer.com/10.1007/s11263-019-01228-7. doi:10.1007/s11263-019-01228-7.