A Pilot Study to Evaluate Success Criteria for Symbiotic AI

Miriana Calvano[1], Antonio Curci[1,2], Paloma Diaz[3] and Rosa Lanzilotti[1]

[1]Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, Italy
[2]Department of Computer Science, University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[3]Department of Computer Science, Universidad Carlos III de Madrid, Leganés, Madrid, Spain

Abstract

Artificial Intelligence (AI) is transforming humans' activities through continuous collaboration. In a context where AI is becoming pervasive, the focus is shifting toward the creation of Symbiotic AI (SAI) systems that support humans rather than replace them allowing them to intervene on system's behavior while providing explanations. This contribution presents a set of *Success Criteria (SC)* that can be used to evaluate SAI systems by assessing whether the requirements of symbiosis are met. This work also presents the protocol for the user study to validate the SC in real-world contexts. To avoid potential issues that can affect the validity of the user study, a pilot survey was performed involving 5 participants in which they were asked to assess the SC based on three metrics (Comprehensibility, Relevance, and Feasibility) with respect to an AI-based system. According to the preliminary results, the SC achieved positive ratings, but the protocol requires refinements since participants often prioritized the usability of the system over the assessment of the SCs.

Keywords

Symbiotic AI, Human-AI Interaction, Evaluation, User Study,

1. Introduction

The spread of Artificial Intelligence (AI) is having a significant impact on individuals' lives, reshaping how they conduct activities and interact with technology. Since AI is being integrated in many contexts, some of which involve taking decisions that affect people's lives, there is a need for systems that enable individuals to understand the reasoning underlying their functioning, allowing them to make informed decisions [1, 2]. Thus, new interaction paradigms are emerging in which humans and AI have a bi-directional collaborative relationship to reach shared objectives [3, 4]. This has brought Human Computer Interaction (HCI) and AI to become deeply intertwined, contributing to the establishment of the field of Human-Centered Artificial Intelligence (HCAI) focused on creating AI systems that reflect human goals, needs and preferences. Within HCAI, Symbiotic Artificial Intelligence (SAI) is also emerging, referring to AI systems that support rather than replace humans, enabling mutual and continuous exchange that benefits both parties over time [5]. Due to the close relationship among AI and humans, and the pervasive nature of current AI systems used to support decisions that affect people in many different ways, SAI systems must comply with legal standards, ensuring the preservation of fundamental rights and values, and the safeguarding of the environment surrounding them (e.g., physical, social, organizational and cultural context). In this regard, the European Union (EU) has released the *Artificial Intelligence Act* (AI Act), which is a legal framework that regulates the design, development, and use of AI systems in the EU to safeguard humans [6, 7]. However, legal compliance is not enough; a human-centric perspective, in which humans are considered in all their dimensions rather than as mere users is also required [6]. AI systems can be enhanced by enabling non-technical users, who are experts in a particular application domain, to participate directly in AI-driven decision-making processes. To this end, End-User Development (EUD) can provide a solution to these problems, allowing people to create, customize, or adapt AI-based systems to their own needs and goals [8]. The interaction

Proceedings of the 10th International Workshop on Cultures of Participation in the Digital Age (CoPDA 2026): Exploring the Relationship between EUD, AI-Assisted Development, and Meta-Design, June 2026, Venice, Italy.
miriana.calvano@uniba.it (M. Calvano); antonio.curci@uniba.it (A. Curci); pdp@inf.uc3m.es (P. Diaz); rosa.lanzilotti@uniba.it (R. Lanzilotti)
0000-0002-9507-9940 (M. Calvano); 0000-0001-6863-872X (A. Curci); 0000-0002-9493-7739 (P. Diaz); 0000-0002-2039-816 (R. Lanzilotti)

mechanisms in SAI are closely related to the Culture of Participation (CoP) as they allow users to be directly involved in the creation of these systems [9, 10]. The first step toward supporting this adaptation is to identify the criteria that determine a SAI system's success in its interaction with humans which, in this paper, are called *Success Criteria* (SC).

Building on already established AI-Act-based principles for SAI defined in the previous work [11], this research presents practical indications, called Success Criteria (SC), to support HCI and AI experts in building SAI systems. In particular, the objective of SC is to support practitioners during the system evaluation by providing verifiable conditions that ensure the requirements of SAI systems. In addition to the set of SC, this contribution introduces a study protocol for evaluating their effectiveness in real-world scenarios and reports preliminary results from a pilot study carried out to verify the feasibility and validity the protocol. The study focuses on a restricted set of SC because the primary aim is to examine human-AI interaction processes, leaving aside those that concern more architectural or technical issues, which will be evaluated in a separate study.

This manuscript is structured as follows: Section 2 provides the motivations behind this research, defining previous research it is based on and outlining the objectives; Section 3 reports the methodology of the study, illustrating the focus, materials, and procedure; Section 4 explores the results of the study and presents the main findings; Section 5 draws the conclusions on this work and presents the future directions.

2. Background and Motivations

The motivation behind this work lies in the need to define a standardized approach to create SAI systems, specifically concerning their evaluation, to ensure that the foundational characteristics of human-AI symbiosis are considered. Building upon the work by Calvano et al., which defines four high-level principles that characterize SAI, we have created more pragmatic and concrete solutions for their evaluation. The four principles are *Automation Level*, *Fairness*, *Protection*, and *Transparency* [11]. They dictate the main requirements for SAI systems, which must guarantee an appropriate level of automation allowing humans to intervene and supervise the system's functioning while being properly transparent and explain the motivations behind their outputs [12, 13]. In this context, a fair and non-discriminatory behavior, protecting humans' privacy, safety, and security should be exhibited [14, 15]. In this perspective, EUD paradigms can play a relevant role, empowering humans to actively shape and tailor the AI behavior to their specific needs shifting the human role from passive supervision to active collaborator [8]. Deriving from the principles, lower-level statements were defined (called guidelines) summarizing their content, as a way to provide a top-down approach to group the SC by topic. The guidelines, which are not the main focus of this study, were decomposed into atomic statements that can be considered the foundation to define the SC providing further details on the characteristics of the elements referenced in the guidelines. Each guideline can be characterized by more than one SC to cover and detail each aspect they refer to. The SC consist of verifiable instructions that make it possible to assess whether the guideline, and the corresponding principle(s), to which they refer is complied with. They specify instructions for compliance, depending on AI technologies, contexts of use, and requirements based on the AI system's risk level. Figure 1 represents the connection among principles, guidelines, and success criteria highlighting the focus of this study.

The SC are formalized in the following way: <Subject> <*obligation*> [<condition>]. *Subject* refers to Human, AI Model (algorithm), AI system (infrastructure where the model operates) or AI output; *Obligation* refers to the requirement or characteristic to be satisfied; *Condition* (optional) adds another level of detail defining the specific circumstance or property for which the SC is satisfied or not. It is optional because while some obligations are self-explanatory, others require further details to clarify the specific criteria for fulfillment.

In total, 16 SC and 5 guidelines are included in this study, which are presented in [16], each considering specific components or aspects of SAI systems mainly focus on the human-AI interaction design, as illustrated in Figure 1. An exemplary set containing three SC is reported below, grouping them by the

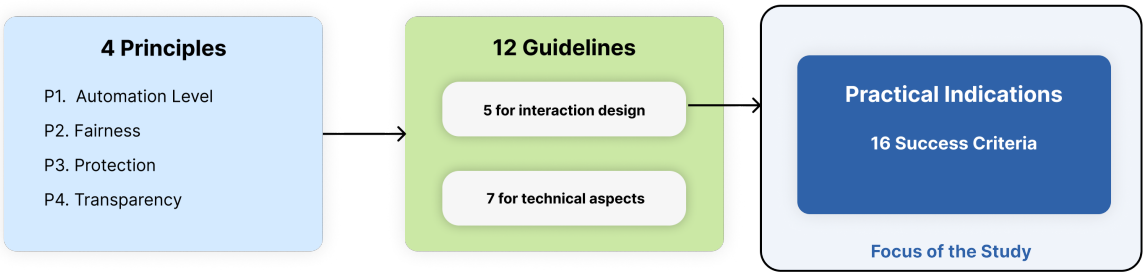

Figure 1: Overview of the process for defining Success Criteria (focus of this study).

guideline they belong to (G1).

Success Criteria for G1 *The human must be enabled to assign proper meaning to the AI model's outputs, whose presentation varies depending on task, processes, and structure to eventually modify its behavior (Transparency, Automation Level).* The SC deriving from it are:

- SC 1.1: The AI system *shows the output with the clear explanation of its meaning and relevance to the task* being understandable to humans (i.e., comprehensible and well-written text and/or images that clearly shows an artifact).
- SC 1.2: The AI system *provides an explanation which must match the context of use, being consistent with real-world scenarios.*
- SC 1.3: The AI system *provides an explanation which must be intended for the specified category of users that the system is intended for.*

To use the SC in real-world contexts as a useful aid for practitioners, their applicability and consistency must be verified. Therefore, the next objective is to investigate whether the proposed SC can be employed to evaluate SAI and, consequently, to assess whether they can be translated into concrete requirements and functionalities of these systems. Considering the above mentioned motivations and objectives, the following Research Question (RQ) was identified:

RQ: Are the Success Criteria effective when evaluating SAI systems?

In this regard, a user study will be conducted to validate and refine the SC investigating whether the system allows user's active participation to customize, intervene and change the behavior of the system. Before performing the study, a pilot experience was carried out to refine the protocol. This pilot study, described in the next sections, will help address potential issues that could compromise the validity of the final user study and, in addition, provides preliminary insights into the quality of the SC.

3. Methodology

The methodology was tested in a pilot study, involving 3 researchers and 2 university students with expertise in HCI and AI, to address potential issues before the extensive study is conducted.

Although the target audience of the systems employed in the study is medical professionals, the goal of this research is to investigate the effectiveness of the SC, which go beyond the specific case and can be applicable to other domains by practitioners. The study protocol was approved by the ethical committee of the University of Bari Aldo Moro with the following identifier: "CER19BFFDE8204".

3.1. Materials

The study draws on various resources, which are described in the following paragraphs.

3.1.1. Identification of Set of Success Criteria to Evaluate

The focus is directed toward a limited set of 5 guidelines and their corresponding SC, with particular emphasis on the human-AI interaction process. Accordingly, this work considers only those SC derived from the interaction-focused guidelines being mappable to User Interfaces (UIs), which are reported in the previous section (Section 2). These goals were used to formulate the following two criteria for the selection of SC: (1) they must be mappable to a UI; (2) they must explicitly mention a design or evaluation method, technique, and/or approach. To provide participants with a broader understanding of the study's context and to ensure an informed evaluation, each guideline is accompanied by the corresponding SC and an example UI that integrates its concepts and description.

3.1.2. Questionnaire

To evaluate the SC, a questionnaire was employed during the study. It contains questions concerning participants' expertise and employment, an introductory presentation on SAI, examples of applying the SC to AI systems, and a set of questions to evaluate the considered SC. The SC are grouped by the guideline to which they refer to, and the corresponding task. These are evaluated using three metrics—defined ad hoc for this study—Comprehensibility, Relevance, and Feasibility, each rated on a 5-point Likert scale. These metrics were previously defined in our previous work, but they were modified accordingly to make them suitable for the SC [17]. Specifically:

(1) *Comprehensibility*. The SC is easy to understand and conveys its message clearly without ambiguities.

(2) *Relevance*. The SC has a significant impact on establishing a symbiotic relationship while not undermining the system's functionality.

(3) *Feasibility*. The SC can be applied to a system's UIs and functionalities.

To gather further insights, for each set of SC, participants were also asked to provide comments on any potential issues encountered as well as suggestions for improvement in a separate question for each guideline.

3.1.3. Interactive Prototypes of AI-based Systems

The user study involves two systems in the medical field that are AI-based and were built following with the objective of reaching human-AI symbiosis, fostering human control through no-code interventions and ensuring that users can modify them according to their preferences. The first is a web application that classifies cells to detect illness from a nasal cytological preparation, based on a digital image acquired directly from a microscope. The second supports neuroscientists in the detection of brain tumors. The system is powered by a multimodal model that classifies 2D grayscale MRI scans and enables clinicians to review and correct diagnoses, thereby improving the model and ensuring that humans are always kept *in-the-loop*, especially in high-risk scenarios. The two of them have different purposes in the context of the study: the first is used in the introductory section of the questionnaire to accompany guidelines and SC with a concrete example of UIs that reflects its content; the second, is used for the actual assessment of the SC based on five tasks that participants were asked to complete.

3.1.4. Tasks

Since participants were not familiar with the system nor with the application domain, they needed to be guided in exploring its functionalities and relevant UIs in order to recognize and evaluate the SC. A total of 5 tasks were defined, each mapped to a single guideline and to the corresponding SC. When participants successfully completed each task, they were asked to fill out the questionnaire mentioned in Section 3.1.2.

(T1) Explore the explanation generated by the system with respect to the latest diagnosis.

(T2) Correct the diagnosis of the system by indicating that the brain is not ill.

(T3) Check the results of the last diagnosis.

(T4) Explore any diagnosis that had been previously modified and revert the results.

(T5) Select any diagnosis and check the system's confidence level

3.2. Procedure

The sessions were divided into two main parts: *Ice-Breakers* and *Success Criteria Evaluation*. During the *Ice-Breakers*, participants were welcomed and informed of the study's objectives. To provide an overall understanding of the study context, the introductory section of the questionnaire presented an overview of SAI, including its principles, guidelines, and SC. Then, if necessary, participants could ask questions to clarify any doubts or issues. Each participant was informed that the data collected in the study cannot be traced back to their identity. Afterwards, the SC were presented, with examples on the UIs belonging to the first system in order to let participants familiarize with a SAI system. After this introduction, the *Success Criteria Evaluation* began, and participants were asked to evaluate them by performing the five tasks on the second AI-based system. The choice to employ two different AI-based systems was motivated by the need to minimise as much as possible biases in the evaluation of SC. For each task, they evaluated the SC by indicating whether its indications were applied in the elements of the UI, assigning a score ranging from 1 to 5 to each metric. To guide participants during the whole study, they were provided with a document that contains all the basic notions about SAI and metrics definitions that they could consult whenever needed.

4. Preliminary Findings of the Pilot User Study

Considering the results gathered from the pilot study, an analysis of comments, inserted by participants for each group of SC, revealed a possible trend that could led to the refinement of the protocol of the extensive user study. Specifically, participants primarily focused their feedback on the system's interaction mechanisms and UIs rather than on the application, content and/or clarity of the specific SC. For instance, several comments addressed aspects concerning the explanations provided by the AI system and the utility of the confidence level of the generated output. While these insights could be valuable for the overall development of the system, they provide limited information regarding the refinement of the SC, representing a potential limitation for the extensive study. The tendency of participants to focus on the prototype's usability instead of SC, suggests the need to modify the presentation of contents in the questionnaire along with improving the training of participants before filling it.

Overall, the feedback from the pilot study was positive, with most criteria receiving high scores (around mean = 4) across all metrics. The SC 5.2 ("The AI system is not provided with humans' personal data as input unless relevant for its foundational functioning") received the best scores across the three metrics (Comprehensibility mean = 4.2; Feasibility mean = 4; Relevance mean = 4.4). Instead, the SC 5.6 ("The AI system embeds ethical guidelines into reconfiguration workflows (e.g., via checklists or rule-based constraints") received the lowest scores concerning Comprehensibility and Feasibility (mean = 3). This could be due to its high level of abstraction and overlapping with the SC 5.5 ("The AI model can be reconfigured whether an ethical compliance test — which verifies that the system aligns with established ethical guidelines and relevant regulations (e.g. AI Act, GDPR) — is passed before deployment."). This result could suggest a potential integration of the SC 5.6 with the SC 5.5 since they refer to strictly connected aspects. In addition, the SC 5.3 ("Humans provide consistent input types with respect to the requirements of the AI system to avoid wrong outputs") resulted to be the less relevant (Relevance mean = 3) for participants; this was not an expected outcome since it is important

to ensure that users insert relevant information for the system's purpose avoiding to provide personal data which should not used to train the AI model, unless exceptional circumstances, to avoid biases and any form of damage. A possible motivation behind this could be the fact that participants were not expert developers in the medical domain, thus underestimating the relevance of this indication.

In the end, being a pilot study these results are not representative; thus, the future extensive study will allow to obtain a broader set of results and richer feedback. Specifically, a more representative sample will be selected by involving individuals with an higher level of experience in HCI and AI. In addition, the protocol will involve an introductory session with more complete indications along with a tutorial of how the activity must be carried out with the objective to show them the relevant aspects to analyze. Another improvement that can be applied is adding reminders in the questionnaire to recall the objectives of the study.

5. Conclusions

Creating SAI systems is an activity that requires detailed and pragmatic guidance to ensure that all the important elements are implemented in the system, allowing humans to intervene on the AI system. The symbiotic paradigm allows also non-technical users to be more in control of the system having an active role in AI-aided decision-making leading. This contribution presents a set of SC, which are practical indications that guide practitioners in creating AI systems that embody these concepts. To evaluate their Comprehensibility, Feasibility, and Relevance, it is intended to conduct a user study which represents a first step toward refining and validating them. To avoid potential issues during the extensive user study, a pilot study was conducted involving researchers and students with expertise in HCI and AI. The preliminary results show that the three metrics attained high values for each SC. Nevertheless, there is a significant limitation since participants focused more on AI system's usability rather than SC.

As future work, the protocol of the user study will be refined to overcome this limitation along with SC, which will also be applied and validated in a real-context AI-based systems. It will be also interesting to further explore similar solutions that can be employed to create AI systems, such as the guidebook defined by Google [1], to expand the current version of SC.

Acknowledgments

This research 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(s) have not employed any Generative AI tools for writing the article.

References

[1] M. H. Jarrahi, Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making, Business Horizons 61 (2018) 577–586. URL: https://linkinghub.elsevier.com/retrieve/pii/S0007681318300387. doi:10.1016/j.bushor.2018.03.007.

[2] F. M. Calisto, J. Fernandes, M. Morais, C. Santiago, J. M. Abrantes, N. Nunes, J. C. Nascimento, Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis, in: Proceedings of the 2023 CHI conference on human factors in computing sys-

[1] https://pair.withgoogle.com/guidebook/

tems, Chi '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 20. URL: https://doi.org/10.1145/3544548.3580682. doi:10.1145/3544548.3580682, place: Hamburg, Germany tex.articleno: 13.

[3] J. C. R. Licklider, Man-Computer Symbiosis, IRE Transactions on Human Factors in Electronics HFE-1 (1960) 4–11. URL: https://ieeexplore.ieee.org/document/4503259. doi:10.1109/THFE2.1960.4503259.

[4] G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Designing and Evaluating Human-Centred AI Systems: Best-Practices from a Multidisciplinary View, in: D. Glowacka, C. Santoro, Z. Xiao (Eds.), Joint Proceedings of the ACM IUI Workshops 2025, volume Vol-3957, CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 170–178.

[5] 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.

[6] European Parliament, Council 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 (Artificial Intelligence Act), 2024.

[7] R. J. Neuwirth, Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA), Computer Law & Security Review 48 (2023) 105798. doi:10.1016/j.clsr.2023.105798.

[8] A. Esposito, M. Calvano, A. Curci, G. Desolda, R. Lanzilotti, C. Lorusso, A. Piccinno, End-User Development for Artificial Intelligence: A Systematic Literature Review, in: L. D. Spano, A. Schmidt, C. Santoro, S. Stumpf (Eds.), End-User Development, volume 13917, Springer Nature Switzerland, 2023, pp. 19–34. URL: https://link.springer.com/10.1007/978-3-031-34433-6_2. doi:10.1007/978-3-031-34433-6_2.

[9] B. R. Barricelli, D. Fogli, A. Locoro, EUDability: A new construct at the intersection of End-User Development and Computational Thinking, Journal of Systems and Software 195 (2023) 111516. URL: https://linkinghub.elsevier.com/retrieve/pii/S0164121222001923. doi:10.1016/j.jss.2022.111516.

[10] D. Tetteroo, S. S. Feger, D. Fogli, A. Mørch, A. Piccinno, M. Utterberg Modén, Participatory design & end-user development: Building bridges, in: Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction, NordiCHI '24 Adjunct, Association for Computing Machinery, New York, NY, USA, 2024, p. 3. URL: https://doi.org/10.1145/3677045.3685466. doi:10.1145/3677045.3685466.

[11] M. Calvano, A. Curci, G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Building Symbiotic Artificial Intelligence: Reviewing the AI Act for a Human-Centred, Principle-Based Framework, Minds and Machines 36 (2025) 1. URL: https://link.springer.com/10.1007/s11023-025-09753-w. doi:10.1007/s11023-025-09753-w.

[12] M. R. Endsley, Supporting Human-AI Teams:Transparency, explainability, and situation awareness, Computers in Human Behavior 140 (2023) 107574. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563222003946. doi:10.1016/j.chb.2022.107574.

[13] P. Kieseberg, E. Weippl, A. M. Tjoa, F. Cabitza, A. Campagner, A. Holzinger, Controllable AI - An Alternative to Trustworthiness in Complex AI Systems?, in: A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, volume 14065, Springer Nature Switzerland, Cham, 2023, pp. 1–12. doi:10.1007/978-3-031-40837-3_1.

[14] M. D. Mueck, A. E. B. On, S. Du Boispean, Upcoming European Regulations on Artificial Intelligence and Cybersecurity, IEEE Communications Magazine 61 (2023) 98–102. URL: https://ieeexplore.ieee.org/document/10061649/. doi:10.1109/MCOM.004.2200612.

[15] M. Wörsdörfer, Mitigating the adverse effects of AI with the European Union's artificial intelligence

act: Hype or hope?, Global Business and Organizational Excellence (2023) joe.22238. doi:10.1002/joe.22238.

[16] M. Calvano, A. Curci, P. Díaz, R. Lanzilotti, A. Piccinno, Success Criteria for Evaluating Symbiotic AI Systems, in: Proceedings of the 2026 International Conference on Advanced Visual Interfaces (AVI 2026), ACM, Venice, Italy, 2026, p. 3. doi:10.1145/3811427.3811515.

[17] M. Calvano, A. Curci, R. Lanzilotti, A. Piccinno, Guiding designers and developers in creating symbiotic ai systems, in: Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter, CHItaly '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3750069.3750095. doi:10.1145/3750069.3750095.