Towards synergistic Human-AI collaboration in Hybrid Decision-Making Systems

Clara Punzi[1,2,3⋆], Mattia Setzu[2,3], Roberto Pellungrini[2,3], Fosca Giannotti[2,3], and Dino Pedreschi[1,3]

[1] University of Pisa, Largo Bruno Pontecorvo, 3, 56127 Pisa, Italy
[2] Scuola Normale Superiore, P.za dei Cavalieri, 7, 56126 Pisa, Italy
[3] KDD Lab, ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy

Abstract. A growing body of interdisciplinary literature indicates that human decision-making processes can be enhanced by Artificial Intelligence (AI). Nevertheless, the use of AI in critical domains has also raised significant concerns regarding its final users, those affected by the undertaken decisions, and the broader society. Consequently, recent studies are shifting their focus towards the development of human-centered frameworks that facilitate a synergistic human-machine collaboration while upholding ethical and legal standards. In this work, we present a taxonomy for hybrid decision-making systems to classify systems according to the type of interaction that occurs between human and artificial intelligence. Furthermore, we identify gaps in the current body of literature and suggest potential directions for future research.

Keywords: Hybrid decision-making systems · Human-machine collaboration · Evaluative AI · Socratic AI · Interactive XAI.

1 Introduction

Artificial Intelligence (AI) can improve decision-making in several ways, particularly by accelerating automated processes and enhancing predictive performance. In both low and high-stakes scenarios, the adoption of black-box models to aid human decision-makers is a prevalent trend. However, these models provide very limited interpretability and interactivity, which raises concerns regarding, among others, transparency, robustness, and fairness. This is especially problematic in high-stakes scenarios, where the absence of the aforementioned properties can lead to severe consequences on human well-being [30].

Establishing *hybrid* systems where humans and AIs synergistically collaborate is crucial to tackle these issues. The overarching goal of such hybrid systems is to leverage the strengths of both humans and AIs to overcome the limitations of both [1]. Hybrid Decision-Making Systems (HDMS) consist of agents with conceptually distinct natures, human and mechanical, that can collaborate in a plethora of ways. However, it is essential to recognize that machines should

⋆ Corresponding author: clara.punzi@sns.it

only serve as auxiliary tools, with humans retaining complete *control* and *agency* throughout the entire decision-making process, as explicitly outlined in the latest ethical and legal documents, notably, the Ethics Guidelines for Trustworthy AI [11] and the European AI Act proposal (Art. 14) [10]. To accomplish this goal, human decision-makers should be equipped with all the necessary tools to comprehend, engage with, and supervise black-box models. State-of-the-art frameworks primarily offer single AI advice along with some kind of explanation derived with the tools of eXplainable AI (XAI) [13]. However, their effectiveness has been called into question as they still do not disclose human agency in its entirety and do they do not always align with the cognitive processes employed by human decision-makers [5, 37]. do not properly align with the human way of reasoning [5, 37]. Our stance is in favor of embracing a framework that considers decision support systems as support tools [6] capable of triggering an evaluative approach towards all plausible options and judgments on a case-by-case basis [26]. We contend that this approach has the potential to enhance individuals' ability to make more informed and less biased hypothesis-based decisions.

Proposals for implementing a synergistic human-AI collaborative setting are not novel. Indeed, numerous solutions have been put forth across multiple disciplines and application domains to address a wide range of challenges and needs [1]. This extended abstract provides an overview of the ongoing research conducted by the authors, which aims to ascertain the state-of-the-art of this field, assess the level of knowledge that has been established, and identify gaps as well as potential avenues for future research. In particular, we put forth taxonomy that encompasses three distinct paradigms for categorising works in the area. These paradigms are determined by the varying degrees of interaction between humans and AI, as well as the level of human agency and control that characterises them. While existing literature has focused on examining certain types of hybrid systems (e.g., [23, 36]) or approaching the subject from non-technical angles (e.g., cognitive [3] or legal viewpoints [4]), our objective is to construct a taxonomy that captures the algorithmic characteristics of hybrid systems.

2 Paradigms of Hybrid Decision-Making Systems

Paradigm 1: Human oversight over algorithms

The most simple and straightforward way of integrating humans in the loop of decision-making systems is through algorithm oversight. In this scenario, machine and human agents operate autonomously, the former executing a specific task and the latter monitoring its execution and deciding whether to accept or reject the AI's output. The purpose of algorithm oversight is to identify any potential malfunctions that may remain undetected in the absence of human supervision. Examples of such failures include dataset shifts (i.e., failures caused by shifts in data distribution), and uncertain predictions (e.g., due to outliers or elevated complexity of the decision context).

This simple paradigm presents several weaknesses, which are to be found in the inherent (and natural) fallibility of overseers in controlling complex techno-

logical systems [18]. Consequently, issues of trust calibration frequently emerge, wherein individuals may exhibit either algorithmic aversion, when they underrely on the AI agent, or algorithm appreciation, when instead they excessively rely upon it. Furthermore, overseers may be prone to overlook questionable or blatantly erroneous algorithmic results due to a flawed understanding or assessment of such outcomes, or due to a reliance on arbitrarily chosen factors suggested by the AI [9, 19]. These concerns are likely to be exacerbated in fairness-related tasks, where personal judgments and pre-existing stereotypes towards marginalized groups may lead to biased supervision of the AI system and bias amplification [14, 22].

Enhancing human oversight via eXplainable AI. To address some of the aforementioned weaknesses, human overseers may be assisted with additional artifacts to allow them to better understand the reasoning of the machine. Explainable Artificial Intelligence (XAI) [13] represents an essential move towards achieving effective synergistic human-machine collaboration. XAI is a subfield of AI that focuses on investigating methods for explicating the rationale underlying the decision-making processes of black-box models in a manner that is comprehensible to humans. XAI can help mitigate some of the shortcomings of the *human oversight* paradigm by facilitating end-users understanding, trust, and efficient management of AI support systems. Notably, Wang et al. [34] devised a theoretical framework in which XAI techniques are integrated to serve two purposes: first, to support human reasoning in line with scientific methods, and second, to reduce decision errors due to cognitive biases resulting from heuristic interferences [17].

Paradigm 2: Humans and Socratic machines

Acknowledging and evaluating the *uncertainty* associated with various solutions is a crucial aspect of decision-making. This is because a final decision can be viewed as the one that minimizes ambiguity with respect to one's intended objectives and beliefs. In the context of HDMS, this translates to accurate uncertainty estimates of the AI computations. Specifically, AI systems can be designed to facilitate the comparison of multiple alternatives along with their uncertainty estimate and machine-generated justification, thus fostering a human-centered perspective to HDMS that more closely resembles natural cognitive processes [20]. Numerous technical solutions have been suggested to implement "Socratic" algorithms that are trained to refrain from revealing their decisions when they recognize that their level of confidence is insufficient, or their performance is inferior in comparison to the one a human could possibly attain. Socratic AI encompasses several families of techniques, including learning to reject [7, 36], selective classification [12], and learning to defer [24, 27]. In particular, the last approach differs from the first two in that it learns to abstain from making predictions based not only on the characteristics of the AI agent but also on the predictive behavior of a human expert, which is represented by additional information in the training data.

Notably, in current Socratic machines, the action of deferral is initiated by the AI system, while the human expert is expected to blindly accept all predictions for the samples over which the AI decides not to abstain. Similarly, the AI system is not designed to receive any feedback from the human during the decision process. In other words, the typical setting is that of a hybrid system where agents work independently of each other. This represents the first huge limitation of Socratic AI, as a truly synergistic and interactive collaboration calls instead for the introduction of *bidirectional* communication that harnesses the potential of both human and AI capabilities. Several other limitations have been pointed out as well [23], including data availability, label scarcity, high computational demands, and limited applicability to real scenarios.

Paradigm 3: Human-AI collaboration

The next natural step in HDMS involves a two-way collaboration where human agents are not relegated to the role of mere executors or overseers but are able to engage in direct interaction with the machine. Here, the primary challenge is establishing proficient communication between humans and machines, whereby both entities engage in a reciprocal learning process. Interaction demands that both humans and machines elucidate their rationale to others, finally enhancing the overall efficacy of the system. The Machine Theory of Mind [29] posits that the core element of a synergistic collaboration is the ability to actively shape the mental model of other agents through repeated communication. Ideally, humans and AI would then attain a shared representation of the whole system, thus aligning knowledge and goals [35], while understanding each other's limitations. However, it should be noted that such aspirations have not been proven achievable with current or foreseen seen technology.

From a technical perspective, interactive HDMSs are defined by several dimensions, such as communication language, time of interaction, and learning cost, which are balanced differently. For instance, in eXploratory Interactive Learning [33, 32], the model requests corrections for pairs of machine-generated decision labels and explanations, so that the human can correct the predicted label and provide an *embeddable* explanation (e.g., an adjustment of feature importance outputs). In other works, the AI system additionally shows its own reasoning (e.g., via rule-based encoded domain logic); this allows the user to provide feedback by either confirming the correctness of such rules [2] or by editing them [15], possibly without even the need to retrain the original model [8, 31].

3 Discussion and future directions

Due to rising societal interest and expanding legal requirements, research is advancing toward frameworks that enable synergistic human-AI collaboration in high-stake decision-making. The proposed taxonomy of HDMS revealed some research issues that remain unresolved. Firstly, there is a need to design interpretable models that facilitate the end-users to engage cognitively and assume

control over the decision-making process. Recent proposals move either towards a co-design methodology [28] (prototyping-testing-redesigning) of explainable AI techniques and user interfaces or towards conversational approaches [16, 25] that utilize dialogue interfaces to present users with diverse explanation formats in a gradual manner, thereby stimulating appropriate cognitive processes. Secondly, it is yet unclear what is the best and most effective mechanism to facilitate bidirectional communication and interaction between humans and AI. The existing body of literature pertaining to Socratic AI may potentially benefit from the latest findings in the field of XAI. Moreover, an exploration of the interplay between paradigm 1 empowered by XAI and paradigm 2 would be a compelling area of inquiry. Third, cognitive theories for decision-making, such as appropriate trust [21] and dual process modeling [17], are not yet properly taken into account in the development of HDMS. Lastly, to validate HDMS, it is necessary to establish human trial designs and co-design strategies [28] that actively incorporate the context of use into the decision-making process.

Acknowledgements

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, and by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI". It has been realised also thanks to "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu) , G.A.No.871042 and by NextGenerationEU - National Recovery and Resilience Plan, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR000001 3 - Notice n. 3264 of 12/28/2021.

References

1. Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., Welling, M.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (2020). https://doi.org/10.1109/MC.2020.2996587
2. Alkan, O., Wei, D., Mattetti, M., Nair, R., Daly, E., Saha, D.: FROTE: feedback rule-driven oversampling for editing models. In: Marculescu, D., Chi, Y., Wu, C. (eds.) Proceedings of Machine Learning and Systems 2022, MLSys 2022, Santa Clara, CA, USA, August 29 - September 1, 2022 (2022)
3. Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E.: Beyond accuracy: The role of mental models in human-AI team performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **7**, 2–11 (Oct 2019). https://doi.org/10.1609/hcomp.v7i1.5285

4. Binns, R., Veale, M.: Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. International Data Privacy Law **11**(4), 319–332 (10 2021). https://doi.org/10.1093/idpl/ipab020, https://doi.org/10.1093/idpl/ipab020
5. Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G.E., Pastore, M.C., Sconfienza, L.M., Folgado, D., Barandas, M., Gamboa, H.: Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine **138**, 102506 (Apr 2023). https://doi.org/10.1016/j.artmed.2023.102506
6. Cabitza, F., Natali, C.: Open, multiple, adjunct. decision support at the time of relational AI. In: HHAI2022: Augmenting Human Intellect. IOS Press (Sep 2022). https://doi.org/10.3233/faia220204, https://doi.org/10.3233/faia220204
7. Cortes, C., DeSalvo, G., Mohri, M.: Learning with rejection. In: International Conference on Algorithmic Learning Theory. pp. 67–82. Springer (2016), https://cs.nyu.edu/ mohri/pub/rej.pdf
8. Elgohary, A., Meek, C., Richardson, M., Fourney, A., Ramos, G., Awadallah, A.H.: NL-EDIT: Correcting semantic parse errors through natural language interaction. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5599–5610. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.444, https://aclanthology.org/2021.naacl-main.444
9. Englich, B., Mussweiler, T., Strack, F.: Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. Personality and Social Psychology Bulletin **32**(2), 188–200 (2006). https://doi.org/https://doi.org/10.1177/0146167205282152
10. European Commission: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (2021), COM(2021) 206 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 [accessed 15 June 2023]
11. European Commission and Directorate-General for Communications Networks, Content and Technology: Ethics guidelines for trustworthy AI (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
12. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
13. Giannotti, F., Naretto, F., Bodria, F.: Explainable for trustworthy AI. In: Human-Centered Artificial Intelligence, pp. 175–195. Springer International Publishing (2023). https://doi.org/10.1007/978-3-031-24349-3_10
14. Grgić-Hlača, N., Lima, G., Weller, A., Redmiles, E.M.: Dimensions of diversity in human perceptions of algorithmic fairness. In: Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022. pp. 21:1–21:12. ACM (2022). https://doi.org/10.1145/3551624.3555306
15. Guo, L., Daly, E.M., Alkan, O., Mattetti, M., Cornec, O., Knijnenburg, B.: Building trust in interactive machine learning via user contributed interpretable rules. In: 27th International Conference on Intelligent User Interfaces. ACM (Mar 2022). https://doi.org/10.1145/3490099.3511111, https://doi.org/10.1145/3490099.3511111

16. Jentzsch, S.F., Höhn, S., Hochgeschwender, N.: Conversational interfaces for explainable ai: a human-centred approach. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1. pp. 77–92. Springer (2019)
17. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus & Giroux, New York, NY (Apr 2013)
18. Koulu, R.: Proceduralizing control and discretion: Human oversight in artificial intelligence policy. Maastricht Journal of European and Comparative Law **27**(6), 720–735 (2020). https://doi.org/https://doi.org/10.1177/1023263X20978649
19. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 29–38 (2019). https://doi.org/https://doi.org/10.1145/3287560.3287590
20. Le, T., Miller, T., Singh, R., Sonenberg, L.: Explaining model confidence using counterfactuals. Proceedings of the AAAI Conference on Artificial Intelligence **37**(10), 11856–11864 (Jun 2023). https://doi.org/10.1609/aaai.v37i10.26399
21. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society **46**(1), 50–80 (Jan 2004). https://doi.org/10.1518/hfes.46.1.50_30392
22. Lee, M.K.: Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society **5**(1), 2053951718756684 (2018). https://doi.org/https://doi.org/10.1177/2053951718756684
23. Leitão, D., Saleiro, P., Figueiredo, M.A.T., Bizarro, P.: Human-ai collaboration in decision-making: Beyond learning to defer (2022). https://doi.org/10.48550/ARXIV.2206.13202
24. Madras, D., Pitassi, T., Zemel, R.: Predict responsibly: Improving fairness and accuracy by learning to defer. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
25. Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a grounded dialog model for explainable artificial intelligence. arXiv preprint arXiv:1806.08055 (2018)
26. Miller, T.: Explainable ai is dead, long live explainable ai! hypothesis-driven decision support (2023). https://doi.org/10.48550/ARXIV.2302.12389
27. Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., Sontag, D.: Who should predict? exact algorithms for learning to defer to humans. In: Ruiz, F., Dy, J., van de Meent, J.W. (eds.) Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 10520–10545. PMLR (25–27 Apr 2023)
28. Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., Rinzivillo, S.: Co-design of human-centered, explainable ai for clinical decision support. ACM Trans. Interact. Intell. Syst. (mar 2023). https://doi.org/10.1145/3587271
29. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.M.A., Botvinick, M.: Machine theory of mind. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4218–4227. PMLR (10–15 Jul 2018)
30. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (May 2019). https://doi.org/10.1038/s42256-019-0048-x

31. Tandon, N., Madaan, A., Clark, P., Yang, Y.: Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 339–352. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.findings-naacl.26, https://aclanthology.org/2022.findings-naacl.26
32. Teso, S., Öznur Alkan, Stammer, W., Daly, E.: Leveraging explanations in interactive machine learning: An overview. Frontiers in Artificial Intelligence **6** (Feb 2023). https://doi.org/10.3389/frai.2023.1066049
33. Teso, S., Kersting, K.: Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM (Jan 2019). https://doi.org/10.1145/3306618.3314293
34. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM (May 2019). https://doi.org/10.1145/3290605.3300831
35. Yang, S.C.H., Folke, T., Shafto, P.: The inner loop of collective human-machine intelligence. Topics in Cognitive Science (Feb 2023). https://doi.org/10.1111/tops.12642
36. Zhang, X.Y., Xie, G.S., Li, X., Mei, T., Liu, C.L.: A survey on learning to reject. Proceedings of the IEEE **111**(2), 185–215 (2023). https://doi.org/10.1109/JPROC.2023.3238024
37. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM (Jan 2020). https://doi.org/10.1145/3351095.3372852