Multi-Agent Evaluation of XAI for Human-AI Trust and Cooperative Adaptation

Francesco Musicco[1]

[1]Politecnico di Bari, Via E. Orabona 4, 70125 Bari, Italy

Abstract

Explainable Artificial Intelligence (XAI) is crucial for ensuring transparency and accountability in high-impact decision-making. However, most existing approaches evaluate explanations through technical metrics, implicitly assuming a monolithic user. In real socio-technical contexts, explanations are interpreted by heterogeneous stakeholders with distinct objectives and skills, leading to interpretative misalignments and decision-making conflicts that remain largely unmodeled.

To address this gap, this doctoral research integrates Multi-Agent Systems (MAS) and reinforcement learning concepts into XAI evaluation. The project pursues two main objectives: (i) developing a MAS framework to simulate interactions among heterogeneous stakeholders and identify interpretative misalignments; and (ii) designing adaptive mechanisms to iteratively optimize explanatory strategies based on collective metrics, such as interpretative alignment and conflict reduction.

Operating exclusively at the explanation generation layer without altering the underlying predictive model, this framework reconceptualizes explainability as a dynamic, relational process. By combining multi-agent simulation with empirical validation, the research aims to systematically improve the robustness and clarity of AI explanations in complex, multi-stakeholder environments.

Keywords

Explainable AI, Multi-Agent Systems, Trust, Reinforcement Learning

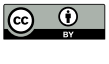
1. Context and Motivation

Artificial Intelligence (AI) systems are increasingly used in high-impact decision-making domains such as healthcare, finance, and the public sector. In these contexts, predictive accuracy alone is insufficient: decisions must also be transparent, auditable, and understandable to human actors involved in oversight and responsibility allocation. For this reason, Explainable Artificial Intelligence (XAI) has emerged as a central paradigm for supporting accountable and human-centered use of AI systems [1, 2]. Despite significant progress in both post-hoc and ante-hoc explanation methods, many XAI approaches are still designed and evaluated assuming a single end user. However, in real-world socio-technical environments, explanations are interpreted by heterogeneous stakeholders, such as domain experts, decision-makers, auditors, or end users, characterized by different objectives, levels of expertise, and decision constraints [3, 4]. As a consequence, the same explanation may generate divergent interpretations, misaligned trust assessments, or conflicting decisions. This situation introduces a structural challenge for XAI: the quality of explanations is determined by the interplay between the technical faithfulness to the model and the ability to sustain coherent interpretation and coordination among multiple actors. Explanations, therefore, function as informational artifacts embedded within relational decision processes, where interpretative conflicts and trust asymmetries may emerge [5]. Current evaluation approaches provide limited support for analyzing such dynamics before system deployment. As a result, interpretative misalignments between stakeholders often emerge only during real system use, when the costs of misunderstanding or disagreement may already affect decision processes.

To address this limitation, this research proposes *CooXAI* (Cooperative XAI), a framework designed to analyze and optimize explanatory strategies in multi-stakeholder environments. The framework models stakeholders as interacting agents with heterogeneous evaluation criteria, enabling controlled

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

f.musicco@phd.poliba.it (F. Musicco)



simulations of how different explanatory configurations influence trust, agreement, and decision-making dynamics. By integrating multi-agent simulation with adaptive mechanisms for explanation refinement, CooXAI aims to support the systematic analysis and improvement of explanatory strategies before deployment, shifting the focus from individual interpretability toward the relational robustness of explanations. Specifically, this research aims to:

1. Develop a multi-agent simulation framework to model the interaction between heterogeneous stakeholders and identify interpretative misalignments in XAI systems
2. Design adaptive mechanisms, inspired by reinforcement learning, to iteratively optimize explanatory strategies with respect to collective metrics such as interpretative alignment, trust consistency, and conflict reduction.

2. Related work

Fundamentals and regulation of XAI. Explainable Artificial Intelligence (XAI) has emerged as a key research area at the intersection of technical, regulatory, and social requirements. Regulatory initiatives such as the European AI Act [6] and the General Data Protection Regulation (GDPR) [7] reinforce the need for transparency, accountability, and verifiability in automated decision-making systems. Despite the growing body of research on interpretability methods, the literature highlights a persistent conceptual fragmentation in the definition of explainability and interpretability [8]. Some approaches prioritize the fidelity of explanations to the internal behavior of the model, while others emphasize cognitive comprehensibility and usability for human decision makers [9]. This tension between technical faithfulness and human interpretability has led to the development of heterogeneous explanatory methods that are often difficult to compare and evaluate systematically. As a result, the effectiveness of explanations remains difficult to assess consistently across different application contexts.

Empirical studies further suggest that the presence of explanations does not automatically improve human–AI collaboration. Experimental evidence shows that explanations may even increase inappropriate reliance on model predictions, leading users to accept incorrect recommendations when explanations appear persuasive [10]. These findings indicate that the value of explanations cannot be assessed solely in terms of transparency, but must also consider their influence on human judgment and decision behavior.

Explanation evaluation and methodological limitations. A central issue in the XAI literature concerns the evaluation of explanations. Despite the large number of interpretability methods proposed in recent years, there is currently no universally accepted framework for determining the quality of an explanatory method. Most technical evaluation metrics—such as fidelity, stability, or sparsity—measure the extent to which explanations reflect the internal behavior of the predictive model, but they do not necessarily capture whether explanations are understandable or useful for human decision-making [11]. To address this gap, recent work emphasizes the need for a multidimensional assessment that integrates these technical properties with human-centric dimensions, such as comprehensibility and decision-making utility [12]. These dimensions are interdependent and may even conflict with each other, making single-metric evaluation insufficient for capturing the full complexity of explanatory performance. In addition, the literature distinguishes between epistemic satisfaction and substantive satisfaction. An explanation may appear convincing or intuitive to users (epistemic satisfaction) while not accurately reflecting the underlying reasoning process of the model (substantive satisfaction). This distinction highlights the risk of conflating communicative clarity with structural validity when evaluating explanations [2]. Another methodological limitation concerns the narrow focus of many evaluation protocols on the interaction between a single user and an AI system. In real-world deployment scenarios, AI systems are rarely interpreted in isolation but are embedded in organizational settings where multiple actors interact with the same model outputs. Under these conditions, the effectiveness of an explanation may depend not only on its individual comprehensibility but also on how it is interpreted and negotiated across different perspectives [13]. These limitations motivate the development of evaluation approaches capable of capturing richer interpretative dynamics.

Multi-stakeholder approaches and social simulation. To address these limitations, recent research has begun to examine explainability within broader socio-organizational contexts. Concepts such as Social Transparency and Extended Human-Centered AI emphasize the importance of accountability, traceability, and role differentiation within AI-assisted decision-making processes [14]. In this perspective, explanations act as mediating constructs that support communication and coordination between actors with different objectives and competencies. Parallel to this line of research, advances in generative agents and role-playing simulations have demonstrated the potential of Large Language Models (LLMs) to represent and simulate complex social behaviors [15]. Frameworks such as CAMEL [16] and multi-agent debate architectures show that interaction between language-model-based agents can produce emergent dynamics of cooperation, conflict, and iterative refinement of responses [17].

In particular, structured debate between multiple language models has been shown to improve reasoning accuracy and factual consistency by enabling iterative critique and revision processes [18]. These interaction patterns highlight how dialogic processes between agents can surface inconsistencies and refine collective reasoning. Beyond simulating individual behaviors, recent work also suggests that LLMs can approximate the responses of broader human populations when conditioned on demographic or contextual information. Studies on algorithmic fidelity demonstrate that language models can reproduce patterns of opinions and attitudes associated with specific social groups, enabling the simulation of heterogeneous human perspectives in computational experiments [19]. These findings open new possibilities for using language models as proxies for diverse stakeholder populations in exploratory analyses of socio-technical systems. However, most existing simulation frameworks are primarily designed to explore the reasoning capabilities of LLMs rather than to systematically evaluate the interpretability of AI model explanations. In particular, current approaches rarely incorporate structured metrics to analyze how different actors interpret the same explanation or how disagreements emerge between perspectives. Consequently, the integration of social simulation with explainability evaluation remains an open research challenge.

Multi-Agent Systems and Reinforcement Learning. Multi-Agent Systems (MAS) provide a formal framework for modeling environments in which multiple autonomous entities interact cooperatively or competitively. Within this paradigm, Multi-Agent Reinforcement Learning (MARL) studies how agents learn strategies through reward signals derived from environmental feedback and interaction with other agents [20]. Recent work combining LLMs with multi-agent architectures explores deliberative interaction patterns such as critique, negotiation, and argumentative exchange between agents [21, 22]. Through iterative comparison and cross-review mechanisms, these systems can improve the consistency and reasoning quality of generated responses. In particular, structured debate between multiple language models has been shown to improve factual accuracy and reasoning performance by enabling agents to critique and revise each other's intermediate conclusions [18]. Several studies also highlight the limitations of purely prompt-based multi-agent interaction. Without explicit learning mechanisms, interactions between pre-trained agents may not lead to stable collaborative strategies or consistent performance improvements [23]. Multi-Agent Reinforcement Learning addresses this limitation by introducing optimization mechanisms based on collaborative rewards and influence-aware signals, enabling agents to learn coordinated behaviors that improve collective outcomes. In parallel, the field of Explainable Reinforcement Learning (XRL) investigates methods for making reinforcement learning policies interpretable through techniques such as reward decomposition, policy visualization, and contrastive explanations [24]. Despite these advances, the literature has not yet developed a systematic framework that integrates multi-agent simulation and reinforcement learning to dynamically evaluate and adapt explanations of AI systems based on collective metrics such as shared trust, interpretative alignment, or conflict reduction. The integration of role modeling, interactive dynamics, and explanation optimization, therefore, remains largely unexplored and motivates the development of new frameworks capable of combining social simulation with explanation evaluation.

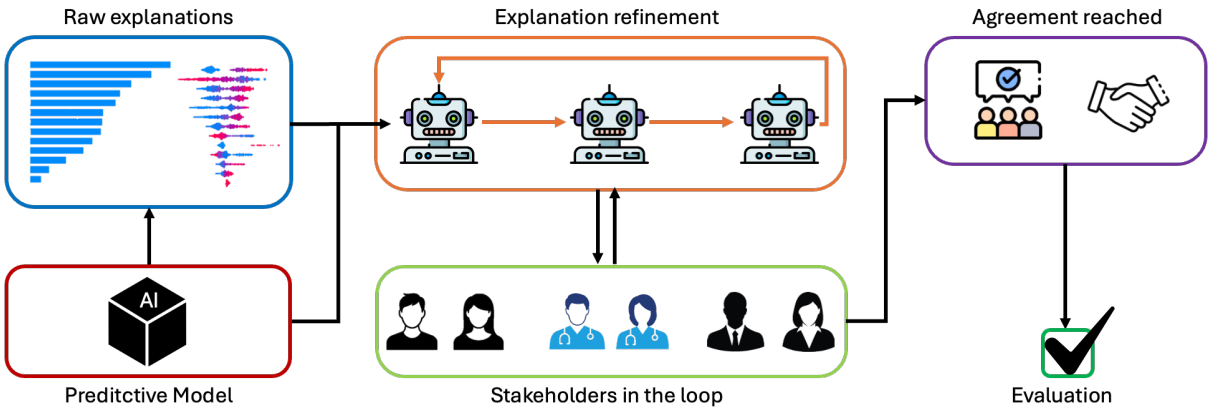


Figure 1: High-level architecture of the proposed CooXAI framework. A predictive model generates a prediction and an associated explanation, which is subsequently processed in a multi-agent environment representing heterogeneous stakeholder roles. Each agent evaluates the explanation from its own perspective, producing role-dependent judgments that enable the analysis of interpretative alignment, disagreement, and ambiguity across stakeholders. These observations support the collective assessment of the explanation and the adaptation of explanatory strategies. The framework also incorporates a stakeholder-in-the-loop mechanism, so that explanations that are not considered sufficiently clear, relevant, or satisfactory for human stakeholders are sent back for further refinement. The process iterates until an acceptable degree of interpretative alignment is achieved among stakeholders.

3. Research questions and methodology

This work proposes *CooXAI*, a framework designed to analyze and adapt AI explanations in multi-stakeholder environments. Figure 1 illustrates its high-level architecture. A predictive model produces a decision together with the corresponding explanation. Rather than modifying the predictive model itself, the framework operates on the explanation layer, processing it within a multi-agent environment where agents represent heterogeneous stakeholder roles. Each agent interprets the explanation according to its own decision criteria, expertise, and informational needs, producing individual assessments that contribute to a collective evaluation of the decision. This simulated interaction enables the observation of interpretative dynamics such as disagreement, ambiguity, or convergence between stakeholder perspectives. These signals are then used to iteratively refine explanatory strategies in order to improve their robustness in multi-stakeholder contexts. To structure this process, the methodology is organized into four interconnected phases corresponding to the following research questions:

- **RQ1 (Stakeholder modeling):** How can heterogeneous stakeholders be computationally modeled to accurately simulate their distinct interpretative behaviors, expertise levels, and evaluation criteria when interacting with AI explanations?
- **RQ2 (Simulation of interpretative dynamics):** How can multi-agent simulations be leveraged to systematically identify interpretative misalignments, ambiguities, and points of conflict across different stakeholder profiles before system deployment?
- **RQ3 (Adaptation of explanatory strategies):** How can explanatory strategies be iteratively adapted, based on simulated collective feedback, to reduce interpretative conflicts and improve cross-stakeholder robustness?
- **RQ4 (Empirical validation):** To what extent do simulation-driven adaptations improve actual human interpretative agreement, and how reliably do simulated dynamics mirror real-world multi-stakeholder interactions?

These questions are addressed through a four-phase methodology encompassing stakeholder modeling, simulation, adaptation, and empirical validation.

Phase 1 – Stakeholder modeling (RQ1)

The first phase focuses on the computational representation of heterogeneous stakeholders as agents representing functional categories of actors (e.g., clinicians, reviewers, end users), characterized by different decision priorities, levels of expertise, and evaluation criteria [13]. Stakeholders are modeled through a hybrid architecture composed of two complementary components. The first is a rule-based structural layer that specifies explicit parameters describing the stakeholder profile, such as information preferences, risk tolerance, and criteria for evaluating explanations. These parameters define the structural constraints for each agent and enable a controlled comparison across different stakeholder perspectives. The second component is a generative layer based on LLMs, responsible for linguistic interpretation and contextual judgment generation [15]. While the rule-based layer defines the decision profile of the stakeholder, the LLM component enables agents to produce qualitative reactions, explanations of their judgments, and deliberative reasoning about the presented explanation. This two-level architecture preserves experimental controllability while maintaining the interpretative flexibility necessary to simulate realistic evaluation dynamics. This phase addresses RQ1 by defining a shared computational representation of heterogeneous stakeholders and enabling the controlled simulation of their interpretative behavior.

Phase 2 – Simulation and analysis of misalignments (RQ2)

The second phase develops a simulation environment to analyze how heterogeneous stakeholders interpret the same AI explanation. In real-world settings, explanations are assessed according to different objectives, expertise levels, and decision criteria. For example, in a patient readmission scenario, a clinician may focus on clinical relevance and patient risk, whereas a hospital administrator may give greater weight to resource allocation and fairness considerations. This phase is designed to capture these different interpretative perspectives and to identify potential points of conflict before deployment. Drawing inspiration from recent work on persona-based multi-agent simulations [15], the methodology follows a two-step design. First, simulated agents independently evaluate the explanation according to their assigned profiles, considering dimensions such as decision utility, risk tolerance, information completeness, and perceived clarity. This stage is intended to capture the baseline interpretation of each stakeholder role while limiting the influence of interaction effects. In the second step, these individual assessments are brought into a collective comparison setting in order to analyze interpretative misalignments across stakeholders. The objective is not to impose a fixed deliberative process, but to identify where explanatory elements are interpreted differently across roles. For instance, the same component of an explanation may be considered sufficiently informative by one stakeholder while remaining ambiguous, incomplete, or unconvincing for another. The main outcome of this phase is a structured characterization of agreement, disagreement, and ambiguity across the simulated stakeholder population. By identifying where interpretative conflicts emerge and how they are distributed across roles, the simulation provides the evidence needed to adapt explanatory strategies in the subsequent phase.

Phase 3 – Adaptation of explanatory strategies (RQ3)

The third phase focuses on adapting explanatory strategies on the basis of the friction points identified during the multi-agent simulation. If Phase 2 shows, for instance, that a clinician considers an explanation too abstract while an administrator finds it excessively technical, the framework should be able to revise the explanation in a way that better addresses these different interpretative needs. To this end, explanations are treated as configurable objects that can be modified along two complementary dimensions: *form* (e.g., level of granularity, narrative structure, or presentation format) and *evidentiary content* (e.g., which features, counterfactuals, uncertainty elements, or fairness aspects are emphasized).

The conflicts, agreements, and ambiguities identified in the previous phase are used as structured feedback to guide this iterative adaptation process. Rather than assuming a single, generic user model, the framework considers a multi-stakeholder setting in which explanatory strategies must be evaluated

with respect to heterogeneous perspectives, priorities, and expertise levels. In this sense, the phase is conceptually informed by recent work on preference-based optimization, including approaches such as Direct Preference Optimization (DPO) [25], while extending the underlying intuition from individual preference alignment to the adaptation of explanations across multiple stakeholder profiles.

A central challenge in this process is ensuring that the adapted explanation reflects meaningful cross-stakeholder robustness rather than the artificial dominance of particular perspectives. In collective simulation settings, some judgments may become disproportionately influential, potentially masking minority but important concerns [26]. For this reason, the adaptation process should account for possible imbalance in stakeholder feedback and preserve the contribution of diverse perspectives when revising the explanation. This phase addresses RQ3 by investigating how explanatory strategies can be iteratively adjusted to reduce interpretative misalignments and improve the overall comprehensibility and usefulness of explanations in multi-stakeholder contexts.

Phase 4 – Empirical validation (RQ4)

The fourth phase focuses on the empirical validation of the proposed framework. The objective is to assess whether explanatory strategies adapted through the multi-agent process improve how heterogeneous stakeholders understand and interpret the rationale behind model decisions, compared to baseline explanations generated directly by the predictive model.

The validation is structured as a comparative evaluation between two experimental conditions: (i) baseline explanations produced by the model using standard explainability techniques, and (ii) explanations adapted through the CooXAI framework following the multi-agent simulation process. This comparison allows the analysis of whether simulation-driven adaptation leads to explanations that are more robust in multi-stakeholder interpretative settings. The evaluation combines quantitative and qualitative metrics commonly used in XAI studies. In particular, the assessment considers three complementary dimensions: perceived clarity of the explanation, task performance in decision-support scenarios, and the level of agreement between stakeholders in the interpretation of the explanation. Particular attention is devoted to analyzing variations in inter-stakeholder agreement when interacting with baseline explanations versus explanations refined through the framework. Empirical validation is conducted through controlled user studies involving participants representing different stakeholder roles. Participants are exposed to decision scenarios in which model predictions are accompanied by explanatory outputs, and are asked to interpret, evaluate, and discuss the explanations in relation to their decision-making tasks. This setting enables the observation of how different actors interpret the same explanation and how explanatory strategies influence collective understanding. Through this evaluation protocol, the framework is assessed with respect to three main aspects: (i) whether adapted explanations improve interpretability across heterogeneous stakeholder profiles, (ii) whether they increase interpretative agreement compared to baseline explanations, and (iii) whether the interpretative patterns observed in the multi-agent simulation correspond to those emerging from real user interactions. Phase 4 addresses RQ4 by providing empirical evidence on the effectiveness of the proposed framework and by evaluating the reliability of simulation-based approaches as support tools for the design and assessment of explanatory strategies in socio-technical decision environments.

4. Preliminary results and contributions to date

Although the project is still in its early stages, several prototype components have already been developed that address parts of RQ1 (stakeholder modeling) and RQ2 (analysis of interpretative misalignments).
Multi-agent prototype for narrative explanations. A multi-agent prototype is currently being developed to transform SHAP-based explanations into narrative descriptions adapted to different simulated stakeholders. The pipeline integrates a predictive model, a SHAP contribution extraction module, and LLM-based agents that reformulate these contributions into natural language. Rather than simply translating numerical values into descriptive text, the agents apply narrative strategies such as contextualization, selective emphasis, and profile-dependent reformulation. This enables the analysis

of how the same quantitative explanation may generate different interpretations depending on the cognitive and decision-making priorities attributed to stakeholders.

Cognitive modeling of stakeholders. Alongside the technical prototype, an ongoing theoretical study investigates the cognitive modeling of stakeholders. Drawing on cognitive agent modeling and argumentation theory, this work formalizes agents through explicit decision priorities, risk acceptability thresholds, and recurring argumentative patterns for justification. Collectively, these foundational developments demonstrate the technical feasibility of multi-agent narrative explainability, establishing a structured baseline for the framework's subsequent simulation, adaptation, and empirical validation phases.

5. Next steps and expected contribution

Methodological roadmap. The next stages of the research extend the current prototype along the four research questions.

First, stakeholder modeling (RQ1) will be further developed by integrating elements from argumentation theory and cognitive agent modeling in order to represent decision profiles through explicit evaluation criteria and recurring argumentative patterns.

Second, the simulation environment (RQ2) will be expanded to support systematic analysis of interpretative misalignments, introducing quantitative metrics to compare autonomous evaluations with interaction-driven dynamics between agents.

Third, mechanisms for adapting explanatory strategies (RQ3) will be implemented through modular procedures for selecting and configuring explanations, and later through optimization techniques inspired by reinforcement learning, with the aim of reducing detected misalignments.

Finally, empirical validation (RQ4) will be conducted through controlled studies involving real stakeholders and comparisons with annotated datasets, in order to assess the correspondence between simulated and human interpretative behavior.

Expected contribution to knowledge. The research is expected to contribute on three complementary levels.

At the theoretical level, it conceptualizes explainability as a relational and multi-stakeholder phenomenon, introducing the notion of relational robustness of explanatory strategies.

At the methodological level, it proposes a multi-agent simulation paradigm as a preliminary socio-technical testing phase for anticipating interpretative misalignments before system deployment.

At the application level, the framework is intended as a modular tool for designing and evaluating explanatory strategies in high-impact decision contexts, supporting XAI evaluation practices that consider both technical fidelity and the stability of interactions between heterogeneous stakeholders.

Declaration on Generative AI

The author used Generative AI tools exclusively for English grammar refinement. The author is solely responsible for all scientific content.

Acknowledgements

This work was partially supported by the following projects: "LIFE: the itaLian system wIde Frailty nEtwork" - local code: T2-AN-12 (CUP D93C22000640001); DEMETRA: "Development of anensemble learning-based, multidimensional sensory impairment score to predict cognitive impairment in an elderly cohort of Southern Italy" (CUP D99J22001970006) Missione 6/componente 2/Investimento: 2.1 "Rafforzamento e potenziamento della ricerca biomedica del SSN", funded by European Commission NextGenerationEU; "IDENTITA - rete Integrata meDiterranea per l'osservazione ed Elaborazione di percorsi di Nutrizione" (CUP B33C22001230003); SPINDOX: Twin netwoRk plAtform in Metaverse via Artificial Intelligence – TramAI"; Codice Pratica 8S3ALJ8 (CUP B96I25000060007).

References

[1] T. Capel, M. Brereton, What is human-centered about human-centered ai? a map of the research landscape, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544548.3580959.

[2] Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[3] What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473. doi:https://doi.org/10.1016/j.artint.2021.103473.

[4] R. R. Hoffman, S. T. Mueller, G. Klein, M. Jalaeian, C. Tate, Explainable ai: roles and stakeholders, desirements and challenges, Frontiers in Computer Science Volume 5 - 2023 (2023).

[5] Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[6] European Parliament, Council of the European Union, Artificial intelligence act, regulation (eu) 2024/1689, Official Journal of the European Union, 2024.

[7] European Parliament, Council of the European Union, General data protection regulation, regulation (eu) 2016/679, Official Journal of the European Union, 2016.

[8] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[9] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: A review on explainable artificial intelligence, Cognitive Computation 16 (2024) 45–74. doi:10.1007/s12559-023-10179-8.

[10] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI), 2021. doi:10.1145/3411764.3445717.

[11] The level of strength of an explanation: A quantitative evaluation technique for post-hoc xai methods, Pattern Recognition 161 (2025) 111221. doi:https://doi.org/10.1016/j.patcog.2024.111221.

[12] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3583558. doi:10.1145/3583558.

[13] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable ai, arXiv preprint arXiv:1810.00184 (2018).

[14] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, J. D. Weisz, Expanding explainability: Towards social transparency in ai systems, Association for Computing Machinery, New York, NY, USA, 2021.

[15] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, Association for Computing Machinery, New York, NY, USA, 2023.

[16] G. Li, H. A. Al Kader Hammoud, H. Itani, D. Khizbullin, B. Ghanem, Camel: communicative agents for "mind" exploration of large language model society, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[17] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, Z. Tu, Encouraging divergent thinking in large language models through multi-agent debate, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024.

[18] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[19] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using language models to simulate human samples, Political Analysis 31 (2023) 337–351. doi:10.1017/pan.2023.2.

[20] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. arXiv:1911.10635.

[21] F. P. Santos, Dynamics of cooperation and conflict in multiagent systems, AAAI Press, 2023.

[22] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, J. Schmidhuber, MetaGPT: Meta programming for a multi-agent collaborative framework, in: The Twelfth International Conference on Learning Representations, 2024.

[23] C. Park, S. Han, X. Guo, A. E. Ozdaglar, K. Zhang, J.-K. Kim, MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025, pp. 30215–30248. doi:10.18653/v1/2025.acl-long.1459.

[24] S. Milani, N. Topin, M. Veloso, F. Fang, Explainable reinforcement learning: A survey and comparative review, ACM Computing Surveys (CSUR) 56 (2024) 1–36. doi:10.1145/3616864.

[25] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference optimization: your language model is secretly a reward model, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[26] L. Xu, L. Wang, H. Xie, M. Zhou, Contextual bandit with herding effects: Algorithms and recommendation applications, Springer-Verlag, Berlin, Heidelberg, 2024.