Towards Computational Models for Reinforcement Learning in Human-AI teams

Francesco Frattolillo[1], Nicolo' Brandizzi[1], Roberto Cipollone[1] and Luca Iocchi[1]

[1]Sapienza University of Rome, Via Ariosto, 25, 00185 Roma RM, Italy

Abstract

In the evolving field of Artificial Intelligence (AI), research is transitioning from focusing on individual autonomous agents to exploring the dynamics of agent teams. This shift entails moving from agents with uniform capabilities (homogeneous) to those exhibiting diverse skills and functions (heterogeneous). At this phase, research on mixed human-AI teams is the natural extension of this evolution, promising to extend the application of AI beyond its traditional, highly controlled environments. However, this advancement introduces new challenges to the learning system, such as trustworthiness and explainability. These qualities are critical in ensuring effective collaboration and decision-making in mixed teams, where mutual cooperation and decentralized control are fundamental. Reinforcement Learning emerges as a flexible learning framework that well adapts to semi-structured environments and interactions, such as those under consideration in this work.

This paper aims to contribute to bridging the gap between Multi-Agent Reinforcement Learning (MARL) and other disciplines that focus on human presence in teams or examine human-AI interactions in depth. We explore how MARL frameworks can be adapted to human-AI teams, highlight some of the necessary modeling choices, discuss key modeling decisions, and highlight the primary challenges and constraints. Our goal is to establish a unified framework for mixed-learning teams, encouraging cross-disciplinary contributions to refine MARL for complex settings.

Keywords

Multi-Agent Systems, Reinforcement Learning, Trust, Computational Modeling, Mixed Human-AI Teams,

1. Introduction

In today's rapidly advancing technological landscape, the integration of AI into our daily lives is becoming increasingly prevalent. As AI applications become more human-centric, the collaboration between humans and AI agents also grow significantly. In collaborative interactions, establishing trust between humans and their AI counterparts is crucial. However, the concept of trust is often not defined in an unambiguous way. Shahrdar et al. [1] highlights the existence of over 300 definitions across various research fields, including notions of measurement, computational models, and human-inspired models.

Trust in human-AI teams has often been evaluated through subjective means, typically via surveys completed by humans post-interaction with AI [2, 3, 4]. While these evaluations capture the subjective nature of trust, rooted in human emotions, beliefs, and experiences, they overlook

MultiTTrust: 2nd Workshop on Multidisciplinary Perspectives on Human-AI Team, Dec 04, 2023, Gothenburg, Sweden

frattolillo@diag.uniroma1.it (F. Frattolillo); brandizzi@diag.uniroma1.it (N. Brandizzi); cipollone@diag.uniroma1.it (R. Cipollone); iocchi@diag.uniroma1.it (L. Iocchi)
0000-0002-2040-3355 (F. Frattolillo); 0000-0002-3191-6623 (N. Brandizzi); 0000-0002-0421-5792 (R. Cipollone); 0000-0001-9057-8946 (L. Iocchi)

the objective and measurable components essential to integrating trust in AI systems. To address this gap, our focus shifts to the explicit computation of these objective quantities. Nevertheless, developing a universal metric for trust in human-AI teams presents significant challenges. Trust inherently depends on multiple factors, including the nature of the autonomous agents, their capabilities, prior expectations of the human teamates, and the current task.

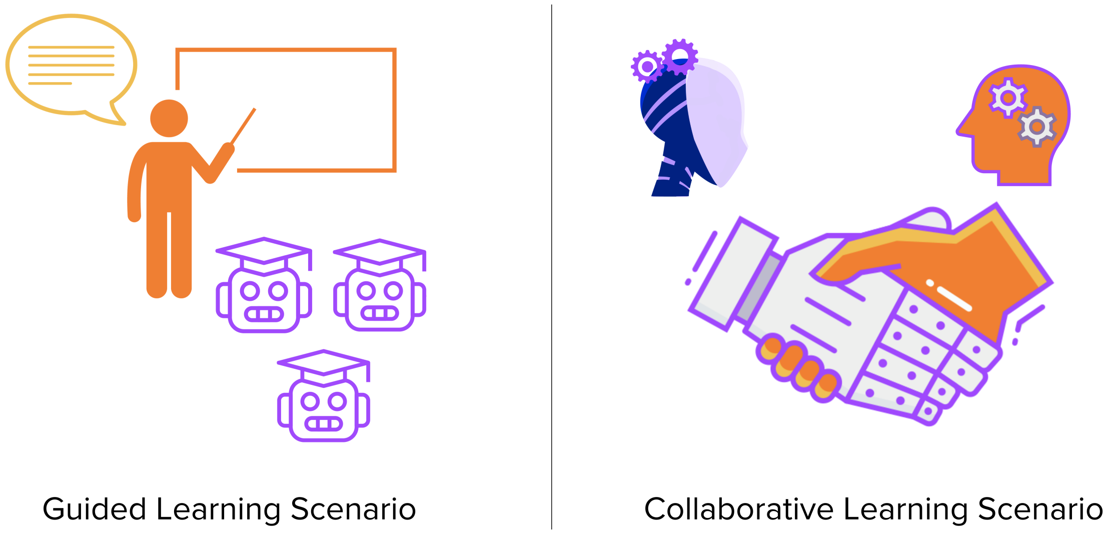
In this paper, we explore trust within the context of a Multi-Agent Reinforcement Learning (MARL). Here, a group of learning agents, both artificial and human, perform different actions at the same time to achieve a common goal. This framework, widely recognized in the Reinforcement Learning community, lays the groundwork for our research. We establish fundamental concepts and terminology to advance the study of trust in mixed learning systems. Additionally, we highlight how specific measures of trust could be defined and assessed within such MARL environments.

2. Multi-Agent Reinforcement learning

In this section, we introduce the basic language that allows to model learning systems with multiple agents. This background knowledge is a prerequisite for the extension for mixed Human-AI teams that we propose in section 3.3. A Markov Game (MG) [5] is a mathematical framework used for modeling multi-agent problems. Formally, a Markov Game is defined as a tuple $\langle N, \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$, where N is the number of players (agents), $\mathcal{S}$ is the set of environment states, shared by all agents, $\mathcal{A}$ is the set of joint actions $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$, where $\mathcal{A}_i$ is the set of actions available to the i-th agent, $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[1] is the transition function returning the probability of the transition from a state to another under the joint action a, and $R : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is the common reward function of all agents. Finally, γ is the discount factor, which is a parameter that quantifies the importance of future rewards compared to immediate rewards. Intuitively, this model defines a joint team state, evolving in a probabilistic way under the joint action of all agents. The team performance is measured via the shared reward function R.

Decentralization and Partial Observability MGs are limited in assuming a fully observable and centralized decision-making environment. However, most real-world scenarios do not satisfy this restrictive assumption, since each autonomous agent must decide its own action independently, given the partial information that is available. To address these constraints, we consider Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) [6]. A Dec-POMDP is defined as a tuple $\langle N, \mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma \rangle$, where N is the number of agents; Ω is the set of joint observations $\Omega = \Omega_1 \times \Omega_2 \times \ldots \Omega_N$, with Ω_i being the observation space for the i-th agent; $O : \mathcal{S} \to \Omega$ is the joint observation function; and $\mathcal{S}, \mathcal{A}, T, R, \gamma$ are defined as in MGs. The key distinction of Dec-POMDPs lies in their accommodation of decentralized decision-making and partial observability. Each agent in a Dec-POMDP operates with its own perspective, limited by individual observation spaces. This feature makes Dec-POMDPs particularly well-suited for modeling the interactions of mixed human-AI team interactions.

[1] $\Delta(\mathcal{X})$ represents a probability distribution over a set of possible values $\mathcal{X}$.



Guided Learning Scenario

Collaborative Learning Scenario

Figure 1: On the left, the *Guided Learning Scenario* depicts humans teaching AI agents, highlighting a unidirectional knowledge flow. On the right, the *Collaborative Learning Scenario* shows bidirectional learning between humans and AI, emphasizing mutual adaptation.

Solutions in Dec-POMDP Environments The goal of Reinforcement Learning algorithms is to learn a function, called *policy*, $\pi : \mathscr{S} \to \mathscr{A}$, mapping states over next actions, that maximizes the expected sum of discounted rewards. The discounted sum of rewards, starting from a time step t is given by the sum:

$$G_t = \sum_{k=t}^{T} \gamma^k R(s_k, a_k)$$

where $0 \leq \gamma < 1$ is the discounted factor. In the specific case of Dec-POMDPs, policies are agent-specific functions that map local observations to available actions. So, the solution of a Dec-POMDP is more properly represented as a set of N policies $\pi_1, \dots, \pi_N$, where each policy is a function $\pi_i : \Omega_i \to \mathscr{A}_i$. Jointly, these agents' policies should maximize the joint expected cumulative reward G_0.

3. Mixed Human-AI teams

In the context of human-AI teams, there are two main scenarios, illustrated in figure 1. In the first case, AI agents learn and adapt based on human instructions; the relationship is predominantly unidirectional: humans teach, and AI learns. We refer to this model as the *Guided Learning Scenario* (GLS). In contrast, the second scenario involves both humans and AI as joint learners. This collaborative approach supports a bi-directional learning process, where both parties contribute to and learn from each other. In this scenario, trust is built on the synergy and mutual adaptation between human and AI capabilities, with each influencing the other's learning curve. This is the *Collaborative Learning Scenario* (CLS).

3.1. Guided Learning Scenario

A review of existing literature on mixed human-AI teams within the RL framework reveals a predominant focus on scenarios where humans act as teachers. In many of these approaches, the human feedback is exactly the reward function R, supplied to the AI agent. In particular, Li et al. [7] presents an extensive survey on human-centered reinforcement learning, identifying three primary approaches: interactive shaping, learning from categorical feedback, and learning from policy feedback. In the *interactive shaping* approach, human observers provide feedback in the form of a shaped reward. For example, Li et al. [7] references "clicker training", initially used for animals, where a clicker sound coupled with food (acting as rewards) shape the animal's behavior. This method was adapted for AI training by Jr. et al. [8] pioneering its use by training an AI agent using reward and punishment in a virtual chat room environment. Contrastingly, *learning from categorical feedback* utilizes categories like positive or negative rewards and punishments [9]. This approach simplifies the feedback mechanism by categorizing it into more intuitive and discrete forms. Lastly, *learning from policy feedback* involves humans directly suggesting the optimal action [10]. Unlike the previous methods, where feedback influences learning indirectly, this approach provides explicit guidance on the actions to be taken, simplifying the decision-making process for the AI. This approach later evolves into techniques to infer the human motivation behind their actions, a concept central to Inverse Reinforcement Learning (IRL). Introduced by Ng and Russell [11], IRL focuses on deducing the reward function guiding observed behavior. This shift from traditional RL, which centers on maximizing predefined rewards, to understanding underlying motivations in IRL, has led to significant advancements. One notable application is Apprenticeship Learning [12], where AI learns complex tasks like driving not by explicit instructions, but by inferring rewards from human behavior. This approach has been further developed in studies on imitation learning [13] and Theory of Mind [14, 15] for AI, underscoring the importance of understanding human intentions and behaviors in mixed human-AI interactions.

3.2. Collaborative Learning Scenario

In the approaches listed in the previous section, humans remain external to the MARL system, as they are not participating to the learning process, but they merely act as teachers, guiding the AI agents. A different approach involves humans as active participants in the learning process alongside AI agents. This perspective shifts the focus to a more integrated and cooperative framework. Here, the MARL setting becomes highly heterogeneous, comprising a mix of human and artificial agents who collaborate and learn together to achieve a shared objective.

Within this collaborative learning context, Cooperative Inverse Reinforcement Learning (CIRL) offers significant insights [16]. In CIRL, both human and AI agents work together to optimize a shared reward function, initially known only to the human. This collaborative approach differs from traditional IRL setups, which typically view humans as isolated optimizers of their own rewards. Optimal CIRL solutions encourage cooperative behaviors like active teaching, learning, and communication from both sides, advancing a stronger alignment of objectives and trust between humans and AI agents. While CIRL has been well-established in theory, its practical applications are still evolving. Research in this area includes experiments where AI agents learn language-driven objectives and adapt to feedback in a manner reminiscent of human learning [17, 18, 19, 20]. However, a notable gap remains in the direct application of these concepts with human participants in non-linguistic contexts. Addressing this gap is crucial not only for enhancing the collaborative learning scenario but also for modeling the trust dynamics within human-AI teams.

3.3. Decentralized RL in human-AI teams

In this section, we extend Dec-POMDPs to human-AI teams. Dec-POMDPs are one of the most commonly used models in multi-agent RL research, and successful integration of this model for mixed teams would be a major step towards the development of a joint learning MARL system. In fact, all participating agents, being either human or autonomous, need to be represented. However, humans are very different from autonomous AI agents, and they may not be simply incorporated into the same setting without modifications. For this reason, we formalize an interaction model that separates humans from AI agents. This allows us to consider a distinct set of actions, states and observations that is specific to each group. To this end, we define the Human-AI Decision process (HADP), as an extension of the classic Dec-POMDP, in which the components that are relative to the human and the AI agents are separated. A HADP is a tuple $\langle \Theta_H, \Theta_A, T, R, O, \gamma, [C], [b] \rangle$. Here, $\Theta_H = \langle N_H, \mathcal{S}_H, \mathcal{A}_H, \Omega_{h_i} \rangle$ represents the elements associated with the human agents (H), and $\Theta_A = \langle N_A, \mathcal{S}_A, \mathcal{A}_A, \Omega_A \rangle$ denotes the ones of artificial agents (A). These represent the state space, the available actions and the observations that are available to the AI agents and the humans. The total number of agents participating in the team is $N = N_H + N_A$. The parameter γ is the discounting, as for Dec-POMDPs. Finally, the variables between square brackets C and b respectively define the communication capabilities and the belief function, specific to each agent. The square bracket suggests that these components are optional. The main distinction between the components in Θ_H and Θ_A is that, while most AI-related elements can be known through direct estimation, the humans' states $\mathcal{S}_H$ and observations Ω_H may not be modeled with the same simplicity. This motivates our choice to separate Ω_H and Ω_A, since the observation functions in a human-AI team can not be merged into a single, joint observation function since humans have no access to the internals of other humans and of the artificial agents. For this reason, it is necessary to either use a shared communication channel C, in order to create a link between AI and human agents, or to approximate their internal state/belief via b_A, by observing their behavior. Related to communications, there are multiple studies that highlight how structured communication can build trust within teams [21, 22, 23], and with the recent progresses of Large Language Model such as ChatGPT [24], Claude [25], and Bard [26], it should be easier to integrate high level communication inside a reinforcement learning scenario.

Related to the belief approximation, there are studies following the Theory of Mind (ToM) principles[2] in which agents learn to synthesize and to use a representation of some key features of other agents' state [27, 28].

[2] Theory of Mind refers to the ability to understand and interpret others' mental states, such as beliefs, desires, and intentions.

4. Trust in Mixed Human-AI teams

After defining the HADP model, this section presents our second contribution, which is to highlight the main components that will necessary appear in any trust metric. We begin by exploring the key variables needed to define a formal trust function within mixed Human-AI teams in Reinforcement Learning scenarios. Trust is a multifaceted concept, influenced by a multitude of factors, making it challenging to distill into a simple, explicit formula. Recognizing the intricate nature of trust, we observe the relevant body of literature with the purpose to contribute to the indentification of some quantifiable components. We start by acknowledging that Trust is a concept strictly correlated to the specific application domain. In order to model Trust in the context of Reinforcement Learning, we should make use of the variables that are available in the RL framework. The one described in section 2 is not the only available. Different RL techniques modify the original framework to account for additional variables related, for example, to communication [29], and to an approximation of other agents' mental state [30], often referred to as belief. Similar to [31], we believe that there are three elements that should be considered when defining a Trust function: a *trustor* X, which is an agent that is currently evaluating the trust; a *trustee* Y, which is an agent or a group of agents that are able through their behavior to build, maintain and improve trust, and a *task* Γ used to evaluate the performance of the trustee. Trust should also be a dynamic function, since the components that influence it, such as the agents, the environment, and even the mental state of agents, are typically dynamic. Following the definition from [32], trust is also strictly related to the concept of risk, and the trustor should be willing to put itself in a vulnerable position with respect to the actions of the trustee. In the context of Reinforcement Learning, this could be implemented through the use of individual reward functions, one for each trustor, that are conditioned on the actions of the trustee. Given these considerations and the model introduced in section 3.3, a trust *function* for an AI agent in a mixed human-AI RL framework should have the following general stucture

$$Trust(X|Y, \Gamma) = f(o^X, a^Y, r, [b^{X\to Y}], [c^{X\to Y}], [c^{Y\to X}])$$

where o^X is the observation of the trustor, a^Y is the action of the trustee, r is the immediate reward, $b^{X\to Y}$ is the current belief that the trustor currently has with respect to the trustee, and $c^{X\to Y}$ and $c^{Y\to X}$ are optional variables that indicate respectively the communication from the trustor to the trustee, and viceversa. We argue that these components are necessary for an effective estimation of trust in a Dec-POMDP. However, further components might be required to account for the peculiarities of the specific environment and task.

5. Conclusions and Future Work

In this paper, we explored Reinforcement Learning in mixed Human-AI teams. In Section 3.3, we proposed a model that accounts for the necessary differences between humans and autonomous agents. We did so, by separating the treatment of the two, and allowing for additional communication or belief construction capabilities. Then, in Section 4, we focus to the problem of definiting objective measures of trust in mixed human-AI learning teams. Instead of defining a specific measure, which would be necessarily domain-specific, we identify the essential component of a flexible trust function. This would necessarily be instantiated in each application with specific modelling choices.

This work is motivated by the growing importance of human-AI collaboration in society. As these interactions become more prevalent, defining and measuring trust in mixed systems becomes increasingly important. Future work should aim to further refine trust definitions into formal mathematical models, leading to a more comprehensive understanding of trust in the context of complex human-AI interactions.

Acknowledgments

This work is supported by the Air Force Office of Scientific Research under award number FA8655-23-1-7257 and PNRR MUR project PE0000013-FAIR.

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

References

[1] S. Shahrdar, L. Menezes, M. Nojoumian, A survey on trust in autonomous systems, Advances in Intelligent Systems and Computing 857 (2019) 368–386. URL: https://link.springer.com/chapter/10.1007/978-3-030-01177-2_27. doi:10.1007/978-3-030-01177-2_27/TABLES/4.

[2] R. E. Yagoda, D. J. Gillan, You want me to trust a robot? the development of a human–robot interaction trust scale, International Journal of Social Robotics 4 (2012) 235–248.

[3] V. Pitardi, H. R. Marriott, Alexa, she's not human but… unveiling the drivers of consumers' trust in voice-based artificial intelligence, Psychology & Marketing 38 (2021) 626–642.

[4] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai, International Journal of Human-Computer Studies 146 (2021) 102551.

[5] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning, Mach. Learn. Proc. 1994 (1994) 157–163. doi:10.1016/B978-1-55860-335-6.50027-1.

[6] D. S. Bernstein, R. Givan, N. Immerman, S. Zilberstein, The Complexity of Decentralized Control of Markov Decision Processes, Mathematics of Operations Research 27 (2002) 819–840. URL: https://pubsonline.informs.org/doi/10.1287/moor.27.4.819.297. doi:10.1287/moor.27.4.819.297.

[7] G. Li, R. Gomez, K. Nakamura, B. He, Human-centered reinforcement learning: A survey, IEEE Transactions on Human-Machine Systems 49 (2019) 337–349. doi:10.1109/THMS.2019.2912447.

[8] C. L. I. Jr., C. R. Shelton, M. J. Kearns, S. Singh, P. Stone, A social reinforcement learning agent, in: E. André, S. Sen, C. Frasson, J. P. Müller (Eds.), Proceedings of the Fifth

International Conference on Autonomous Agents, AGENTS 2001, Montreal, Canada, May 28 - June 1, 2001, ACM, 2001, pp. 377–384. URL: https://doi.org/10.1145/375735.376334. doi:10.1145/375735.376334.

[9] R. T. Loftin, J. MacGlashan, B. Peng, M. E. Taylor, M. L. Littman, J. Huang, D. L. Roberts, A strategy-aware technique for learning behaviors from discrete human feedback, in: C. E. Brodley, P. Stone (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada, AAAI Press, 2014, pp. 937–943. URL: https://doi.org/10.1609/aaai.v28i1.8839. doi:10.1609/AAAI.V28I1.8839.

[10] S. Griffith, K. Subramanian, J. Scholz, C. L. I. Jr., A. L. Thomaz, Policy shaping: Integrating human feedback with reinforcement learning, in: C. J. C. Burges, L. Bottou, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 2625–2633. URL: https://proceedings.neurips.cc/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

[11] A. Y. Ng, S. Russell, Algorithms for inverse reinforcement learning, in: P. Langley (Ed.), Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000, Morgan Kaufmann, 2000, pp. 663–670.

[12] P. Abbeel, A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, in: C. E. Brodley (Ed.), Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, volume 69 of *ACM International Conference Proceeding Series*, ACM, 2004. URL: https://doi.org/10.1145/1015330.1015430. doi:10.1145/1015330.1015430.

[13] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, An algorithmic perspective on imitation learning, Found. Trends Robotics 7 (2018) 1–179. URL: https://doi.org/10.1561/2300000053. doi:10.1561/2300000053.

[14] J. Ruiz-Serra, M. S. Harré, Inverse reinforcement learning as the algorithmic basis for theory of mind: Current methods and open problems, Algorithms 16 (2023) 68. URL: https://doi.org/10.3390/a16020068. doi:10.3390/A16020068.

[15] J. Ruiz-Serra, M. S. Harré, Inverse reinforcement learning as the algorithmic basis for theory of mind: Current methods and open problems, Algorithms 16 (2023) 68. URL: https://doi.org/10.3390/a16020068. doi:10.3390/A16020068.

[16] D. Hadfield-Menell, S. Russell, P. Abbeel, A. D. Dragan, Cooperative inverse reinforcement learning, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 3909–3917. URL: https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html.

[17] T. R. Sumers, R. D. Hawkins, M. K. Ho, T. L. Griffiths, D. Hadfield-Menell, Linguistic communication as (inverse) reward design, CoRR abs/2204.05091 (2022). URL: https://doi.org/10.48550/arXiv.2204.05091. doi:10.48550/ARXIV.2204.05091. arXiv:2204.05091.

[18] H. Liu, C. Sferrazza, P. Abbeel, Chain of hindsight aligns language models with feedback, CoRR abs/2302.02676 (2023). URL: https://doi.org/10.48550/arXiv.2302.02676.

doi:10.48550/ARXIV.2302.02676. arXiv:2302.02676.

[19] K. Nguyen, D. Misra, R. E. Schapire, M. Dudík, P. Shafto, Interactive learning from activity description, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8096–8108. URL: http://proceedings.mlr.press/v139/nguyen21e.html.

[20] T. R. Sumers, R. D. Hawkins, M. K. Ho, T. L. Griffiths, D. Hadfield-Menell, How to talk so your robot will learn: Instructions, descriptions, and pragmatics, arXiv preprint arXiv:2206.07870 (2022).

[21] S. L. Jarvenpaa, D. E. Leidner, Communication and trust in global virtual teams, Organization Science 10 (1999) 791–815. URL: https://doi.org/10.1287/orsc.10.6.791. doi:10.1287/orsc.10.6.791. arXiv:https://doi.org/10.1287/orsc.10.6.791.

[22] J. R. Allert, S. R. Chatterjee, Corporate communication and trust in leadership, Corporate Communications: An International Journal 2 (1997) 14–21. URL: https://doi.org/10.1108/eb046530. doi:10.1108/eb046530.

[23] K. Boies, J. Fiset, H. Gill, Communication and trust are key: Unlocking the relationship between leadership and team performance and creativity, The leadership quarterly 26 (2015) 1080–1094.

[24] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[25] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional ai: Harmlessness from ai feedback, 2022. arXiv:2212.08073.

[26] J. Manyika, An overview of Bard: an early experiment with generative AI, Technical Report, Technical report, Google AI, 2023.

[27] R. Raileanu, E. Denton, A. Szlam, R. Fergus, Modeling others using oneself in multi-agent reinforcement learning, 2018. arXiv:1802.09640.

[28] H. He, J. Boyd-Graber, K. Kwok, H. Daumé, III, Opponent modeling in deep reinforcement learning, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, 2016, pp. 1804–1813. URL: https://proceedings.mlr.press/v48/he16.html.

[29] C. Zhu, M. Dastani, S. Wang, A survey of multi-agent reinforcement learning with communication, arXiv preprint arXiv:2203.08975 (2022).

[30] J. Jara-Ettinger, Theory of mind as inverse reinforcement learning, Current Opinion in Behavioral Sciences 29 (2019) 105–110. URL: https://www.sciencedirect.com/science/article/pii/S2352154618302055. doi:https://doi.org/10.1016/j.cobeha.2019.04.010, artificial Intelligence.

[31] C. Castelfranchi, R. Falcone, Trust Theory: A Socio-Cognitive and Computational Model, John Wiley & Sons Ltd., Chichester, GBR, 2010.

[32] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (1995) 709–734.