Revisiting Rogers' Paradox in the Context of Human-AI Interaction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Humans learn through individual experimentation and social observation, with different strategies carrying distinct costs and success rates. Rogers' Paradox demonstrated that in simple population simulations, cheap social learning provides no fitness advantage over individual learning alone—a counterintuitive result given centuries of human social learning success. As AI systems increasingly serve as sources of social learning while simultaneously learning from humans, we revisit Rogers' Paradox in the context of human-AI interaction. We extend Rogers' original simulations to examine networks where humans and AI systems learn together about an uncertain world. We propose and evaluate learning strategies across three stakeholder levels: individual humans, AI model builders, and society or regulators. Our analysis examines how these strategies impact the quality of society's collective world model, including potential negative feedback loops where learning from AI may hinder humans' individual learning abilities.

1 Introduction

For centuries, humans have learned about the world in different ways: from each other, and from individually conducting experiments and exploring the world around us. From young children [23] to pioneers like Isaac Newton, Marie Curie and Archimedes, humans have long engaged with the world and each other in many ways like scientists – tinkering with our models of the world [7, 58] and sharing this knowledge to impact society's collective understanding of the world. But conducting experiments yourself – whether exploring out in the world, or thinking really hard to process what you have already observed – often comes with costs. It takes time and energy to think and to explore, and we have fundamental constraints on such resources [44, 24]. Sometimes, it is easier to just build on the behavior or insights from another person; to update your understanding of the world based on what someone else has done [55].

The question of the relative advantage of social versus individual learning among humans was thrown for a loop when Alan Rogers uncovered a seeming paradox: the availability of cheap social learning does not increase the relative fitness of a population compared to a population consisting entirely of individual learners [57]. This has led to many follow-ups exploring what strategies any single learner can employ to mitigate this challenge [21, 42, 15, 16, 50]. Here, we reconsider Rogers' Paradox in the context of today's more powerful and, at the surface, human-compatible AI systems, specifically systems that can engage with language. To date, we can think of these systems as having "socially learned" from us: they have been trained on effectively all of what we humans have written on the web [2]. To an extent, any "world model" or other "understanding of the world" that may or may not have been implicitly learned within one of these models [78, 27, 33, 48, 43, 49, 26, 73] could then arguably have been socially learned from our actions and experiments in the world, that we wrote

Submitted to the NeurIPS 2025 Workshop on Algorithmic Collective Action. Do not distribute.

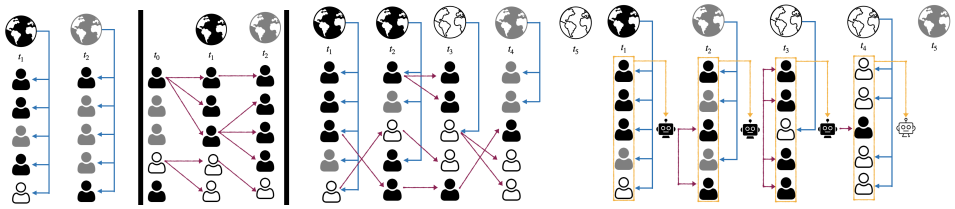

Figure 1: **Traditional Rogers' Paradox.** In the leftmost panel, we depict individual learning (blue arrows). In the second panel, we depict social learning (purple arrows), which is delayed by one timestep. In the third panel, we depict an example of humans oscillating between individual and social learning. Agents are colored by the behavior they adopt in the current timestep. Agents are considered "adapted" if their behavior matches the current state of the world, and have an increased probability of surviving to the next timestep. In the rightmost panel, humans can perform individual learning or learn from the AI system, which reverts to the population mean of the previous timestep.

about in language[1]. The fluency with which these systems can engage with natural language, and therefore us, raises the prospect that now and in the ensuing years, the directionality of learning may flip: humans may increasingly socially learn from these systems [13, 8]. While work has already begun to explore what may unfold if AI systems learn from their own output, or the output of other AI systems [62] — in this work, we focus on possible trajectories from humans and AI systems learning together.

We take preliminary steps to understand a **network of agents**, wherein many humans interact with a single AI system (where the AI system also learns from people), and characterize interconnected ripple effects on a *society* of such learning agents. To start probing possible network effects, we extend the population model introduced in Rogers Paradox with a simulated AI agent: an agent that learns from all other agents. We emphasize that our work then focuses on possible learning *equilibria* induced by a society of different agents, with distinct learning strategies and costs, in a way that could mimic an abstract network model of humans engaging with an AI system. We emphasize, however, that our "AI system" and simulated "human-AI interactions" are highly simplistic, abstract models of agents and their interaction; part of the power, as in much complex systems research, comes from the simulation of many such interactions with simple components [41, 47]. Rogers' Paradox and the follow-up works spawned to understand and mitigate Rogers' Paradox have contributed to a richer conceptual understanding of collective behavior and inspired new empirical work, despite – and sometimes because of – necessarily simplistic simulations [76, 52, 54, 55]. Here, we too hope that revisiting Rogers' Paradox in the context of human-AI interaction will inspire new ways of thinking about and studying possible network propagation effects of AI systems among previously well-characterized networks of human individual and social learners where the world is fundamentally uncertain and dynamic.

This work is structured as follows. We first (Section 2) review Rogers' Paradox and extend the network model to consider an AI system (with different costs) that learns from the entire population at once is introduced into the network. We then (Section 3) consider possible strategies that can be enacted by different actors involved in the interaction and network of interactions: the human, the AI model builders, the interface designers around the moment of interaction, and policymakers writ large. We then turn to a different model of network dynamics (Section 4) wherein interacting with an AI system impacts an agent's ability to individually learn.

2 Integrating Human-AI Interaction in Rogers' Paradox

Rogers' Paradox [57] considers the case where agents in a changing environment try to adapt their individual (simplified) "world models" by either learning about the environment individually or by learning socially from a number of other agents (e.g., their cultural parents). The underlying "true"

[1]Here, we define "world model" as queryable understanding of the world. We refer to AI systems learning a collective world model abstractly (as if there were a single, global "ideal" model). We are not claiming that current large language models have or have not learned such a world model, nor whether such a global model is feasible.

environment, or world, is continually changing; a behavior that may be adaptive in one moment in time may no longer be advantageous if this world changes. The population has some average level of fitness which can be thought of as **the quality of a collective "understanding" of the world** (see Figure 1).

In this world, individual learning is often considered costly and risky: the environment is stochastic, and the agent has some chance of failing to adapt to individual learning. Social learning may be cheaper and uncertainty-reducing, if many other agents have the same strategy then that may provide a signal of the quality of that strategy. However, social learning in a changing environment relies on some other agents having already successfully individually learned the adaptation and as a result, social learning is time-lagged compared to individual learning (providing a potentially outdated "world model" if the environment has changed). Intuitively, the availability of cheap social learning ought to be helpful for improving quality of collective world model by allowing efficient propagation of information; yet, the cheapness creates an incentive for individual learners to become social learners. Accordingly, a relative shift in the proportion of individual learners in a network reduces the availability of timely new information. The fitness of the population at the resulting equilibrium is the same as if there were only individual learners meaning cheap social learning does not improve the collective world model: this is Rogers' paradox [57], which we re-instantiate through simulations, as discussed in Supplement B.

In this work, we extend the simulations to introduce an abstract "AI model" into the network. With the propagation of AI systems like GPT, we increasingly see people using these systems for knowledge retrieval, decision-making, and problem-solving - a clear form of social learning. However, these AI systems themselves are trained on virtually all the text ever produced by humans. Books, text, and the Internet can be thought of as cultural artifacts that reflect our collective world model; accordingly, AI systems can be thought of as simultaneously doing social learning from the entire population at once. The increased availability of these systems makes social learning from the AI system (and by virtue of the way the model is trained, the population therein) very cheap, and the fact that they could be viewed as reflecting a collective world model of the entire population maximizes the uncertainty-reducing effect of social learning; however, we hypothesize that these qualities alone are insufficient to resolve Roger's Paradox. We include more specifics on the network in the Supplement.

3 Exploring Strategies for AI Rogers' Paradox

When there is an AI agent introduced into the network (that learns from us), how can we design for positive learning outcomes across the population? There are multiple parties that can help make human-AI interaction "go right": the human, the developer of the AI model, and the developer of the infrastructure around the model (e.g., the interface builder, even "society" that may change affordances that guide use of an AI system). We consider several strategies that are motivated by the literature on human-AI interaction. For each strategy, we explore how one possible instantiation in our simulation framework may modify the collective "world model" acquisition and quality. We emphasize that our simulations act as a guide for imagining the impact that these strategies may have on *population dynamics* (not just the *individual* interacting with the AI system), as they have in evolutionary biology, anthropology, and other disciplines, extending to thinking about human-AI interaction.

3.1 Human- and Interaction-Centric Strategies

We first consider strategies that can be undertaken by the individual human or addressed in the infrastructure around learning, e.g., by organizations, developers, or regulators.

3.1.1 When Should You Learn from an AI System?

Deciding who to learn from and when is not always an easy task with humans [61, 35]. Deciding when (and when not) to learn from an AI system is an ever more important question as these tools grow more powerful and accessible. In the context of human-AI interaction, there is a burgeoning literature for approaches that encourage the non-use of AI assistance in favor of human judgment [51, 3, 68]. Within the network models around Rogers' Paradox, this equates to humans *choosing* to engage with individual learning over social learning from the AI system. We as authors certainly advocate for critically appraising whether to engage with an AI system. We consider this idea in the context

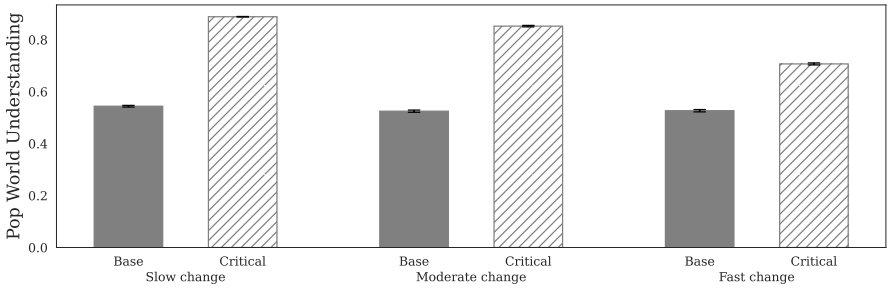

Figure 2: Impact of critical social learning (thatched) from the AI over the baseline learning strategy in the presence of AI (filled). Critical social learning leads to increased population world understanding, across varying rates of world change ($u = 0.01, 0.1, 0.5$); however, critical social learning is a less powerful strategy if the world is changing very rapidly.

of our simulations by assuming that each agent can assess the relative expected utility of a given computation: accounting for their likely world knowledge versus that of the model, and the cost of individually boosting their knowledge versus engaging with the AI system. This appraisal could be done by each individual, but also may come with a cost (e.g., monitoring the quality of the AI system at test-time and assessing one's own abilities).

One could imagine then that humans are told *a priori* about the AI system's ability, quality, or cost and can use this information to decide where they should perform individual learning or social learning, which would correspond to engaging with the AI system. However, when we implement such a strategy – by making the AI system unavailable when the expected adaptation value of learning from it is lower than the expected adaptation value of learning individually – into our network model, we find that the population equilibria of world understanding do not change. This lack of a change to long-term equilibria from such a strategy is in line with other simulations extending the Rogers' Paradox formalism with strategies that turn out to improve individual but not collective fitness [6]; we may imagine that introducing an AI system with different relative costs may change such network equilibria but do not find that is the case. While the research we discuss above fairly definitively identifies the *individual* benefits of this type of strategy, here, we see that it does not translate into improvement in the *collective* world understanding, but it does greatly reduce reliance on the AI system. Future work is well-poised to consider what collective impacts may arise from alternate "who" strategies, e.g., based on model transparency (see Section E).

3.1.2 When Should You Override the Output of an AI System?

After you have decided to engage with an AI system for learning, humans still have *agency to decide* whether or not to uptake the system's output into their thinking. While the decision around *whether* to engage with an AI system does not necessarily impact the equilibria of population world understanding in our current simulations, we next consider the impact of the decision to update one's knowledge of the world upon accessing the model. Any human who uses an AI system for learning will be affronted with a decision on whether they integrate a system's output into their beliefs. We implement this in our simulations with a **critical social learning strategy**: a human can decide to access the AI system, and *based on its output, can decide instead to switch to individually learning instead.* We find that a critical social learning strategy (see Figure 2) yields a *higher* equilibrial state of collective world understanding. Simulation details can be found in Supplement C.

3.2 Model-Centric Strategies

We next explore some strategies that can be applied at the level of the model. a series of questions that warrant consideration to help guide the possible futures we laid out above in a positive direction. We include simulations for several other possible model-centric strategies in the Supplement.

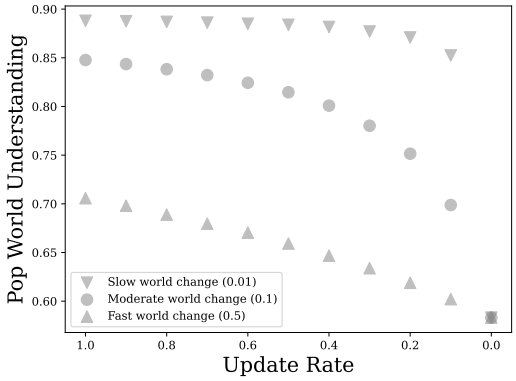

Figure 3: Impact of the update schedule of AI on the collective world understanding. Each dot represents the average population world understanding quality attained with the specific update schedule, in a world of a particular change rate. We consider three different rates of environment (world) change (u). Update rate signifies the probability that the AI will "update" (social learn) on any given timestep.

3.2.1 How Often Should an AI System Update Its Understanding of the World?

While humans can dynamically update their understanding of the world on-the-fly, the process by which an AI system can efficiently update its model remains in question [75]. In practice, however, there are very real costs to AI systems updating its understanding of the world [31, 36, 56]. This raises a question of strategy: in an ever-changing, interconnected world of learners, what is the impact of variable update schedules on collective world understanding? Thus far, we have assumed that the AI system snaps to the population mean on each iteration (t). We next consider the impact of the AI systems' "update schedule". We find in Figure 3 that there is a saturation point at which the frequency of updates has minimal impact on the equilibrium of the population's collective world model understanding (thereby, costs could be saved by a model builder from less frequent updates) that depends on the base rate of change in the environment. Yet, with too infrequent updates, we see deleterious impacts on the equilibria of collective world understanding. These impacts are exacerbated in a base world that is changing more rapidly, rendering learned knowledge and behaviors quickly obsolete, as we see in Figure 3. Simulation details can be found in Supplement C.

4 When Interactions Change Learning Efficacy

Now, imagine if you could at the snap of a finger engage instantly with an expert to learn about the world? What if you never needed to pore over a textbook for hours on your own to learn a new concept – to get stuck and need to start again? While to (hopefully) many, this would not be a particularly fun or fulfilling future: if you can always socially learn for cheap and therefore "avoid" individual learning, your *future ability* to re-engage with individual learning may be substantially weakened. Thus far, we have focused on the impact of various learning strategies in a modified version of the original Rogers Paradox setting wherein a human attempting to improve their understanding of the world can choose to learn individually or socially from another agent (human or AI). However, this choice has no impact on the human other than improving (or failing to improve) that person's understanding of the world. Whichever way you choose to learn, you will always have the same expected learning success. Yet, the ways that you choose to learn *can* impact how successful your future learning may be. This is especially a concern with AI tools or possible "cognitive extenders" [29]. Several works have raised concerns about the impact of AI tools on our cognition and relative self-appraisal: these include algorithm appreciation [46], loafing [32, 60], algorithm aversion [17, 18], algorithmic vigilance [80], de-skilling [53, 22]. We therefore next introduce a slightly richer network of agent interactions into our simulations: *negative feedback upon interaction*.

The simulation framework spawned from Rogers' Paradox offers a fruitful environment to explore the impact of this kind of negative feedback-induced deskilling. We take a step toward exploring this idea by implementing a negative feedback loop: when a human chooses to learn socially from the AI

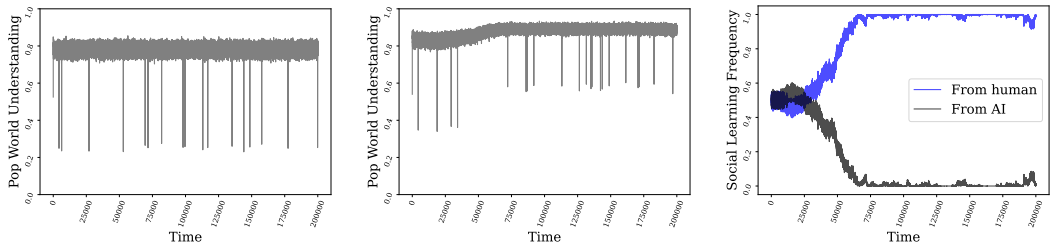

Figure 4: **Left**: Critical social learning with access only to AI and there is negative feedback (learning from AI makes you worse at individual learning). **Center**: Critical social learning with access to both AI and humans (with negative feedback from AI). **Right**: Learners start in the low equilibrium but phase out the AI after a period of time to reach the high equilibrium.

system, the success rate of their own individual learning decreases . To our knowledge, this is a new innovation on top of the base Rogers' Paradox, whereby engaging with another entity in the network change the *learning efficacy* of a given agent. We see in Figure 4 (left) that when there is negative feedback from learning from the AI, the equilibrium collective world understanding dips below that of baseline critical social learning. This makes sense: access to another agent who makes you worse at your own ability to learn (and who too is influenced by your understanding of the world) will run themselves downward to a globally poorer equilibrium. However, in Figure 4 (middle), we see that allowing agents two options for social learning: a more expensive human versus a cheaper AI (which degrades your own ability to learn) can maintain high collective world understanding. As depicted in Figure 4, the population *adapts* to *which source* to learn from socially. Over time, the agents determine that they are better off learning from the agent that does not weaken their own ability to individually learn when they need to. We could imagine more intricate and realistic extensions upon this framework, wherein an agent may shift from negative to positive reinforcement: if so, humans may benefit from a signal (e.g., a nudge [70]) that it is worth learning from the AI to sidestep algorithmic aversion [17]. Details on our negative feedback simulation can be found in Supplement D. We then consider how the choice of method for how the AI system updates its world model effects the population in Supplement C.

5 Conclusion

Rogers' Paradox spawned a rich line of work exploring the relative benefits of individual versus social learning among networks of humans in a dynamic, uncertain world. The introduction of powerful AI systems, which learn from us and which we now may increasingly learn from, opens up new questions about the ripple effects on our collective understanding. Our preliminary network simulations shed light on the importance of critically appraising whether learning from an AI system is sensible – where that appraisal can be made by a human, the AI model builder, or introduced through other scaffolding around the moment of the human-AI interaction (e.g., in the design of interfaces, or regulation).We offer a new extension of these simulations that also models the potential of an AI system impacting the efficacy by which we can in turn, individually learn about the world. However, the simulations we introduce and extend here are necessarily simplistic and only a first step. Future work is well-served to explore the impacts of other modifications to such network models (e.g., introducing multiple AI systems with different costs, or a wider range of possible behaviors in the world). We include a more expansive overview of possible future directions in Supplement E. AI systems are increasingly transforming our understanding of the world and each other. The coming years demand work across multiple levels of abstraction and fields to understand the requirements for, and effects of, deploying evermore capable AI systems into our cultural fabric.

References

[1] K. R. Allen, K. A. Smith, and J. B. Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020. doi: 10.1073/pnas.1912341117. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1912341117`.

[2] J. Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, 2022.

[3] U. Bhatt and H. Sargeant. When Should Algorithms Resign? A Proposal for AI Governance. *Computer*, 57(10):99–103, 2024.

[4] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, and P. e. a. Sattigeri. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, New York, NY, USA, 2021. ISBN 9781450384735.

[5] U. Bhatt, V. Chen, K. M. Collins, P. Kamalaruban, E. Kallina, A. Weller, and A. Talwalkar. Learning personalized decision support policies. *arXiv preprint arXiv:2304.06701*, 2023.

[6] R. Boyd and P. J. Richerson. Why does culture increase human adaptability? *Ethology and sociobiology*, 16(2):125–143, 1995.

[7] N. R. Bramley, B. Zhao, T. Quillien, and C. G. Lucas. Local search and the evolution of world models. *Topics in Cognitive Science*, 2023.

[8] L. Brinkmann, F. Baumann, J.-F. Bonnefon, M. Derex, T. F. Müller, A.-M. Nussberger, A. Czaplicka, A. Acerbi, T. L. Griffiths, J. Henrich, et al. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

[9] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[10] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

[11] A. M. Chen, A. Palacci, N. Vélez, R. D. Hawkins, and S. J. Gershman. A hierarchical bayesian model of adaptive teaching. *Cognitive science*, 48(7):e13477, 2024.

[12] K. M. Collins, U. Bhatt, and A. Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52, 2022.

[13] K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

[14] D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

[15] D. Deffner and R. McElreath. When does selection favor learning from the old? social learning in age-structured populations. *PloS one*, 17(4):e0267204, 2022.

[16] D. Deffner, V. Kleinow, and R. McElreath. Dynamic social learning in temporally and spatially variable environments. *Royal Society open science*, 7(12):200734, 2020.

[17] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[18] B. J. Dietvorst, J. P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3): 1155–1170, 2018.

[19] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[20] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–19, 2021.

[21] M. Enquist, K. Eriksson, and S. Ghirlanda. Critical social learning: a solution to rogers's paradox of nonadaptive culture. *American anthropologist*, 109(4):727–734, 2007.

[22] M. Glickman and T. Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, pages 1–15, 2024.

[23] A. Gopnik. The scientist as child. *Philosophy of science*, 63(4):485–514, 1996.

[24] T. L. Griffiths. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883, 2020.

[25] D. Gunning and D. Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[26] W. Gurnee and M. Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[27] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

[28] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, and P. Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551, 2017. URL `http://arxiv.org/abs/1706.06551`.

[29] J. Hernández-Orallo and K. Vold. Ai extenders: the ethical and societal implications of humans cognitively extended by ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 507–513, 2019.

[30] G. Hinton, O. Vinyals, and J. e. a. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[31] A. Ibrahim, B. Thérien, K. Gupta, M. L. Richter, Q. G. Anthony, E. Belilovsky, T. Lesort, and I. Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=DimPeeCxKO`.

[32] I. Inuwa-Dutse, A. Toniolo, A. Weller, and U. Bhatt. Algorithmic loafing and mitigation strategies in human-ai teams. *Computers in Human Behavior: Artificial Humans*, 1(2):100024, 2023.

[33] A. A. Ivanova, A. Sathe, B. Lipkin, U. Kumar, S. Radkani, T. H. Clark, C. Kauf, J. Hu, R. Pramod, G. Grand, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv e-prints*, pages arXiv–2405, 2024.

[34] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[35] R. L. Kendal, N. J. Boogert, L. Rendell, K. N. Laland, M. Webster, and P. L. Jones. Social learning strategies: Bridge-building between fields. *Trends in cognitive sciences*, 22(7):651–665, 2018.

[36] K. Khetarpal, M. Riemer, I. Rish, and D. Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

[37] H. R. Kirk, A. Whitefield, P. Röttger, A. M. Bean, K. Margatina, R. Mosquera, J. M. Ciro, M. Bartolo, A. Williams, H. He, B. Vidgen, and S. A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=DFr5hteojx`.

[38] J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[39] J. Kleinberg, J. Ludwig, S. Mullainathan, and M. Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024.

[40] H. Kotek, R. Dockum, and D. Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[41] D. C. Krakauer. Unifying complexity science and machine learning. *Frontiers in Complex Systems*, 1:1235202, 2023.

[42] S. Lew-Levy, W. van den Bos, K. Corriveau, N. Dutra, E. Flynn, E. O'Sullivan, S. Pope-Caldwell, B. Rawlings, M. Smolla, J. Xu, et al. Peer learning and cultural evolution. *Child Development Perspectives*, 17(2):97–105, 2023.

[43] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

[44] F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

[45] R. Liu, J. Geng, J. C. Peterson, I. Sucholutsky, and T. L. Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

[46] J. M. Logg, J. A. Minson, and D. A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[47] M. Mitchell. Complex systems: Network thinking. *Artificial intelligence*, 170(18):1194–1212, 2006.

[48] M. Mitchell. Ai's challenge of understanding the world, 2023.

[49] M. Mitchell and D. C. Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[50] E. Miu, N. Gulley, K. N. Laland, and L. Rendell. Flexible learning, rather than inveterate innovation or copying, drives cumulative knowledge gain. *Science advances*, 6(23):eaaz0286, 2020.

[51] H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[52] S. E. Perry, A. Carter, J. G. Foster, S. Nöbel, and M. Smolla. What makes inventions become traditions? *Annual Review of Anthropology*, 51(1):419–436, 2022.

[53] J. Rafner, D. Dellermann, A. Hjorth, D. Verasztó, C. Kampf, W. Mackay, and J. Sherson. Deskilling, upskilling, and reskilling: a case for hybrid intelligence. *Morals & Machines*, 1(2): 24–39, 2022.

[54] L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010.

[55] L. Rendell, L. Fogarty, W. J. Hoppitt, T. J. Morgan, M. M. Webster, and K. N. Laland. Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in cognitive sciences*, 15(2):68–76, 2011.

[56] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023.

[57] A. R. Rogers. Does biology constrain culture? *American anthropologist*, 90(4):819–831, 1988.

[58] J. S. Rule, J. B. Tenenbaum, and S. T. Piantadosi. The child as hacker. *Trends in cognitive sciences*, 24(11):900–915, 2020.

[59] A. Salinas, P. Shah, Y. Huang, R. McCormack, and F. Morstatter. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15, 2023.

[60] S. Saluja, S. Sinha, and S. Goel. Loafing in the era of generative ai. *Organizational Dynamics*, page 101101, 2024.

[61] B. Schwartz. The paradox of choice. *Positive psychology in practice: Promoting human flourishing in work, health, education, and everyday life*, pages 121–138, 2015.

[62] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[63] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[64] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[65] I. Sucholutsky, R. M. Battleday, K. M. Collins, R. Marjieh, J. Peterson, P. Singh, U. Bhatt, N. Jacoby, A. Weller, and T. L. Griffiths. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, pages 2036–2046. PMLR, 2023.

[66] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, and A. B. et al. Getting aligned on representational alignment, 2023.

[67] I. Sucholutsky, K. M. Collins, M. Malaviya, N. Jacoby, W. Liu, T. R. Sumers, M. Korakakis, U. Bhatt, M. Ho, J. B. Tenenbaum, et al. Representational alignment supports effective machine teaching. *arXiv preprint arXiv:2406.04302*, 2024.

[68] S. Swaroop, Z. Buçinca, K. Z. Gajos, and F. Doshi-Velez. Accuracy-time tradeoffs in ai-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 138–154, 2024.

[69] M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024.

[70] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[71] M. Tomasello, A. C. Kruger, and H. H. Ratner. Cultural learning. *Behavioral and Brain Sciences*, 16(3):495–511, 1993. doi: 10.1017/S0140525X0003123X.

[72] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005.

[73] K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan. Evaluating the world model implicit in a generative model. In *Neural Information Processing Systems*, 2024.

[74] G. Vidal. Explanations as programs in probabilistic logic programming. In *International Symposium on Functional and Logic Programming*, pages 205–223. Springer, 2022.

[75] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, and V. K. e. a. Mansinghka. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, pages arXiv–2306, 2023.

[76] C. M. Wu, R. Dale, and R. D. Hawkins. Group coordination catalyzes individual and cultural intelligence. *Open Mind*, 8:1037–1057, 2024.

[77] I. Yanai and M. J. Lercher. It takes two to think. *Nature Biotechnology*, 42(1):18–19, 2024.

[78] I. Yildirim and L. Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 2024.

[79] O. Zaidan, J. Eisner, and C. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL*, 2007.

[80] J. Zerilli, U. Bhatt, and A. Weller. How transparency modulates trust in artificial intelligence. *Patterns*, page 100455, 2022.

[81] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 2023.

A Additional Background on Rogers' Paradox

Consider the following thought experiment (setup adapted from Deffner and McElreath [15] and notation adapted from Enquist et al. [21]). Suppose we have agents ($N = 1000$) in a world that is slowly changing over time, with a fixed probability ($u = 0.01$) that the optimal behavior for succeeding in this world changes at each time step (e.g., due to some weather event or other change to the affordances in the world). Agents in this environment who discover the optimal behavior (consistent with Enquist et al. [21] we call this the "OK" behavior) for a given timestep are "adapted" and have an increased probability of surviving ($s^{OK} := P(\text{survival}|\text{OK}) = 0.93$) compared to non-adapted individuals ($s^{\neg OK} := P(\text{survival}|\text{not OK}) = 0.85$). To discover the current optimal behavior, agents can attempt to learn about the current state of the world at a cost ($c_I = 0.05$) and with some risk of failure (success probability $z_i := P(\text{OK}|\text{individual learning} = 0.66$, $p_i^{OK} := (1 - c_I)z_i = 0.95 * 0.66 = 0.627$. At the end of each timestep, agents survive according to their survival probability, and the environment is replenished back to its original size with new agents. At the end of each timestep after this process, we measure the proportion of agents who are adapted (q^{OK}), and we track this quantity over many timesteps ($T = 200000$) to find the long-run equilibrium fitness of the population: $E[q^{OK}] = p_i^{OK} s^{OK}$[2].

Now suppose that we introduce a second learning strategy into this environment: we allow some agents to learn socially by copying the behaviors they observe from another randomly selected agent. We assume social learning is much cheaper ($c_S = 0$) and more reliable than individual learning, such that an agent is guaranteed to learn the optimal behavior if they observed another agent performing it and the environment has not changed that timestep, $p_s^{OK \to OK} := P(\text{OK}|\text{observed OK}) = P(\text{unchanged})P(\text{copied successfully}) = (1 - u) * 1 = 0.99$. Thus, the probability of becoming adapted when social learning is equal to the proportion of adapted agents in the environment, $p_s^{OK} := (1 - c_S)q^{OK} p_s^{OK \to OK}$). Agents can either be individual or social learners (with proportions of $q_i, q_s = 1 - q_i$ of each type in the environment, respectively), but when new agents are added to the environment at the end of each timestep, they inherit their learning strategy from the surviving agents. Intuitively, we would expect that adding social learning, a cheap and reliable method of learning behaviors, would increase the average fitness; but surprisingly, the equilibrium reached by this network has the *same* average fitness as when the agents only had access to individual learning. This is Rogers' Paradox [57]. We visualize this setup in Figure 1.

While initially unintuitive, Rogers' Paradox can be understood as arising from the *tension* between individual learners having an incentive to become social learners when social learning has higher expected fitness ($E[q_s^{OK}] = E[p_s^{OK}]s^{OK}$) than individual learning ($E[q_i^{OK}] = p_i^{OK} s^{OK}$), and the expected fitness of social learners depending on the number of individual learners in the network.

$$E[q^{OK}] = \frac{p_i^{OK} s^{OK} E[q_i]}{[1 - (1 - c_S)p_s^{OK \to OK} s^{OK} E[1 - q_i]]} \tag{1}$$

We can see in Equation 4 that the expected mean fitness across all agents in this case depends on the proportion of individual learners. The expected mean fitness of just social learners (Equation 5) thus also depends on the proportion of individual learners.

$$E[q_s^{OK}] = \frac{(1 - c_S)p_s^{OK \to OK} s^{OK} p_i^{OK} s^{OK} E[q_i]}{[1 - (1 - c_S)p_s^{OK \to OK} s^{OK} E[1 - q_i]]} \tag{2}$$

As this network evolves over time, if the fitness of social learners is higher than individual learners, the proportion of individual learners will decrease until their propagation rates are equal resulting in the same mean fitness at equilibrium as when only individual learning was available – even though both individual and social learners are present in the network. These results can be validated empirically by running simulations. We run simulations adapted from Enquist et al. [21] and [15] and visualize the results of these baseline simulations in Figure 6 (left) and Supplement Figure 5. We provide implementation details in Supplement B, as well as derivations.

[2] In all simulations we explore in this paper, as discussed in the Supplement, the base $E[q^{OK}] = 0.58$.

A.1 Introducing AI into the Network

We extend our simulations for studying Roger's Paradox to give agents in a slowly changing environment three choices: learn individually (has cost - $c_i = 0.05$, may fail resulting in agent not adapting - $z_i = 0.66$ - with overall success probability $p_i^{OK} = (1 - c_i)z_i = 0.627$), socially from a randomly selected other agent (no cost - $c_s = 0$, succeeds only if the other agent is adapted - $p_s^{OK} = (1 - c_s)q^{OK}p_s^{OK \to OK}$), or socially from an AI system (no cost - $c_{AI} = 0$, succeeds with probability equal to adaptation level of the AI, $p_{AI}^{OK} := q^{OK}$ - see below). At the end of each timestep, **the AI system learns socially from the entire population and matches the corresponding probability distribution of strategies (i.e., the AI's adaptation level is set to the mean adaptation status of the population)**. Agents are operationalized as being individual or social learners, and social learning agents have a propensity for learning from other agents versus from the AI system. Agents who successfully learn the correct strategy become adapted and have a slightly higher survival probability - so every timestep, the population-level propensity, for individual versus social learning and for social learning from agents vs from the AI, is updated. Running these extended simulations we again find that the average population fitness does not increase relative to the individual learning-only case (see Figure 6). In fact, even if we decrease the relative cost of learning from the AI (by increasing c_s, the cost of socially learning from other agents) or increase the relative transmission success probability when learning from the AI (by decreasing $p_s^{OK \to OK}$) the mean population fitness at equilibrium (Equation 3) remains unchanged.

$$E[q^{OK}] = p_i^{OK} s^{OK} E[q_i] + E[p_s^{OK}] s^{OK} E[q_s] + E[p_{AI}^{OK}] s^{OK} E[q_{AI}] \tag{3}$$

This suggests a novel form of Roger's Paradox for the AI age: the widespread availability of cheap AI systems trained on all human data in the world may not, on its own in the long-term, improve our collective world model. In the following section, we consider existing and novel directions in human-AI interaction and foundation model research through the lens of Roger's Paradox to evaluate their potential effect on our ability to leverage AI for improving our collective world model. We include additional details on the simulation environment in Supplement B.

B Additional Details on Base Simulations

We include additional details on our simulations. All code will be open-sourced upon publication at the following repository: https://github.com/collinskatie/ai-rogers-paradox.

Base Network Model

We first set up social learning simulations in the style of the original Rogers' Paradox designs, as discussed in Section 2. We simulate 1000 agents in a slowly-changing, stochastic environment (P of 0.01 that the environment changes at each timestep). Agents who are adapted to the environment at each timestep have a slightly increased chance of continuing to the next timestep (P of not being eliminated 0.93 instead of 0.85), but when the environment changes, all agents become maladapted until they learn the new strategy. We consider two types of agents: individual learners and social learners. At each timestep, depending on their type agents can either attempt to learn individually to try to adapt to the environment, or to learn socially and copy another agent's strategy. Learning individually from the environment has a cost and a risk of failure (P of success = 0.66) but if successful, provides up-to-date information on the environment making the agent adapted. Learning socially from another (randomly selected) agent is free, but the learner only becomes adapted if the teacher was adapted, relying on potentially stale information from a previous round. At the end of each timestep, the population is replenished back up to 1000 agents with new agents descended from the remaining ones with a small mutation rate causing them to flip their learning strategy (P of mutation is 0.005).

We run one simulation where social learning is unavailable and there are only individual learners, and a second simulation as described above where both social and individual learners are present (see Figure 5. We run each simulation for 200,000 timesteps and measure the proportion of adapted agents at each time step. We average the proportion over the final 50,000 steps to get the population fitness (collective world "understanding").

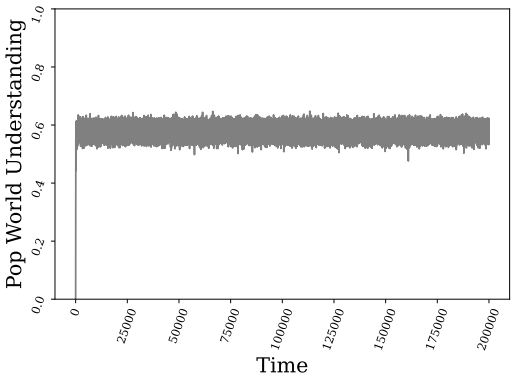

Figure 5: Quality of population world understanding in the original baseline Rogers' Paradox setting: only individual or social learning from other humans allowed. This is identical, in equilibrium, to a network of only individual learners, or a network of humans and an AI social learner.

We also include a derivation of expected fitness, as introduced in Section 2, including the expected mean for all agents:

$$\begin{aligned} E[q^{OK}] &= p_i^{OK} s^{OK} E[q_i] + E[p_s^{OK}] s^{OK} E[q_s] \\ &= p_i^{OK} s^{OK} E[q_i] + (1 - c_S) p_s^{OK \to OK} s^{OK} E[q^{OK}] E[1 - q_i] \\ &= \frac{p_i^{OK} s^{OK} E[q_i]}{[1 - (1 - c_S) p_s^{OK \to OK} s^{OK} E[1 - q_i]]} \end{aligned} \tag{4}$$

And the expectation for just social learners:

$$\begin{aligned} E[q_s^{OK}] &= E[p_s^{OK}] s^{OK} \\ &= (1 - c_S) p_s^{OK \to OK} s^{OK} E[q^{OK}] \\ &= \frac{(1 - c_S) p_s^{OK \to OK} s^{OK} p_i^{OK} s^{OK} E[q_i]}{[1 - (1 - c_S) p_s^{OK \to OK} s^{OK} E[1 - q_i]]} \end{aligned} \tag{5}$$

Introducing AI to the Network

We extend the base simulations to modify how the AI system learns at each step (Figure 6). The AI system either learns by snapping to the mean of the population, or "pays" a cost of individually learning. We vary the cost of individual learning and expected accuracy of the AI system individually learning. We assess the relative gains in the base Rogers' Paradox scenario (when humans either individually or socially learn based on their reliance propensity) versus a population wherein people engage in critical learning. We illustrate additional network dynamics in Figure 7. We use the same parameters controlling world dynamism, human preferences, and human learning quality as in the previous sections.

C Additional Details on Strategy Simulations

When Should You Learn from an AI System?

We assume that the individual learning success probability is known, and performance of AI system (i.e., conditional probability of success when learning from AI assuming the environment has not changed) can be evaluated. We extend the simulation to consider the case where social learning from the AI is only available when the probability of becoming adapted after learning from the AI exceeds the probability of becoming adapted after learning individually. This can be considered a form of

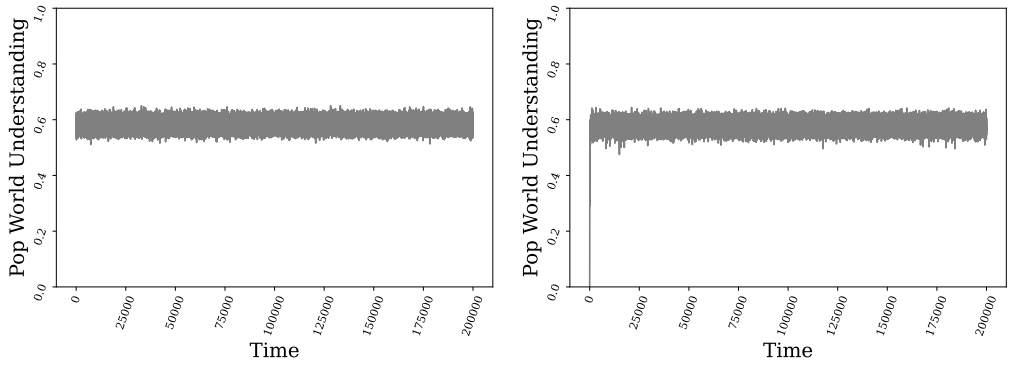

Figure 6: Comparing the collective world understanding over time, in a network where agents can only learn individually and do so at a cost (left) versus a network where agents can learn individually at the same cost or learn socially for free from an AI system that socially learns from all agents in the network (right). Each network attains the same baseline expected collective world understanding: recovering the classic Rogers' Paradox finding, even with an AI node.

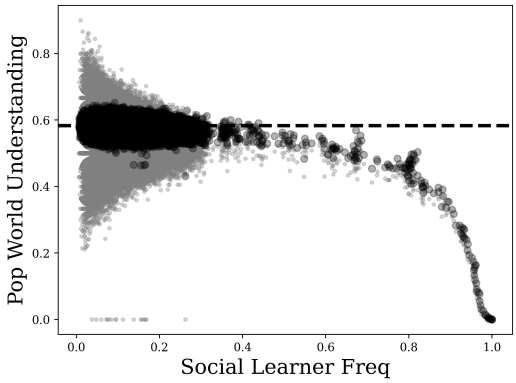

Figure 7: Social learning frequency versus population world understanding for the baseline AI Rogers' Paradox case from Section 2. The dashed line indicates the equilibrium of collective world understanding from a network of purely individual learners. Black indicates mean understanding of all learners; gray indicates mean of just social learners. When expected value of social learning exceeds that of individual learning, individual learners become social learners and vice-versa, resulting in the same equilibrium as in the individual learning case.

conditional social learning where the agent always attempts to social learn, unless it is known that social learning is unlikely to succeed relative to individual learning. Unlike typical conditional social learning strategies, here the agent decides whether to social learn or individual learn *before* seeing the output of the teacher (in this case the AI). This particular strategy has no net impact on the equilibria of collective world understanding, as depicted in Figure 8.

When Should You Override from an AI System?

We next imagine that the agent can tell whether they became adapted after attempting to learn: that is, the agent can *evaluate* how successful a learning interaction was, in line with [21]. We extend the simulation to consider the case where agents can override the outputs of the AI system by learning individually if social learning first fails. All other parameters are the same.

How Often Should an AI System Update Its Model of the World?

Rather than having the AI system socially learn on every iteration, we consider a variable update schedule where the model socially learns with probability u on each iteration. We sweep over possible

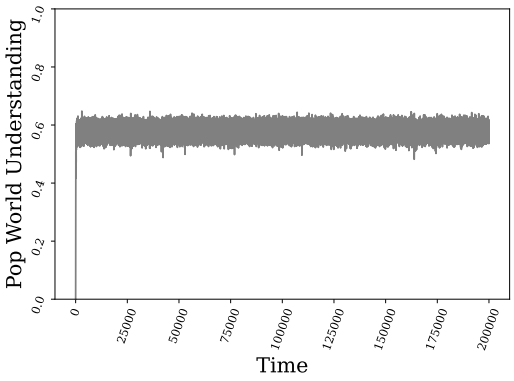

Figure 8: Impact of permitting an agent to decide whether to learn individually or socially from the AI; there is no net impact on collective world understanding.

update schedules u and assume the population has implemented critical social learning. All other simulation parameters are the same.

How Should an AI System Update Its Understanding of the World: Socially or Individually?

While many popular AI systems today (e.g., several large language models) may be viewed as having socially learned from us, the quality of such an AI system's "understanding" of the world is then based on the knowledge of the individual humans in the system: the advantage to the population comes from the AI system's ability to consolidate and provide access to this knowledge at a cheap cost to other humans, not "new" knowledge per say. To go beyond what humans know to bring our collective "world model" closer to the actual world, we may also consider an AI system which updates based on its individual interactions with the world. Many researchers in AI have, and are, pursuing alternate training schemes, wherein AI systems learn from non-human generated data (e.g., distilling information from from other AI models [30], engaging with explicit simulators of the world [1, 28], or relying on self-play [63]). Recent approaches have begun examining test-time compute scaling as an additional way to improve AI performance (albeit at a fairly steep cost) by letting the AI 'think' for extended periods of time before responding [64]. Increasing discussion around "agentic" AI also opens up the possibilities that AI systems individually learn about the world by taking action in that world, for instance, iteratively checking code against a compiler, searching the web, or taking physical action.

Whether an AI system has learned socially (from us) or individually (from self-play or other actions), we may learn from such an AI system through the same mechanisms. We explore this idea by considering an AI agent in our population network which can learn individually, with some cost relative to social learning. But in practice, just because an agent can individually learn does not mean that such learning will be successful. We take first steps to explore the impact of varied individual learning success rates and varied individual learning costs on population dynamics in Figure 9. We see that when the AI system has a low cost to individually learning and a high success rate when pursuing individual learning – the population's net collective understanding of the world can improve substantially, regardless of whether the population (or infrastructure around users' interactions with the AI) involves critical evaluation of use. However, we see that when the AI is often *unsuccessful* with individual learning (i.e., it comes to a bad conclusion about the world) and particularly when costs to individual learning are low such that the model is frequently individual learning, then population world understanding is hampered, unless the human agents have not already shifted to critically appraising the model's output. These data further drive home the intuition that if the AI's individual learning is unsuccessful, it is all the more crucial that an individual human, or infrastructure around their access, e.g., including transparency, critically assess whether it is worth uptaking the output of an AI system. Notably, we see that the *composition* of strategies is especially crucial when the AI system either has a higher cost to individual learning, or lower success rate: it remains important that the population engages in some form of critical appraisal on whether or not to override the output of the AI system. Exploring the relative benefits of compositions of human-

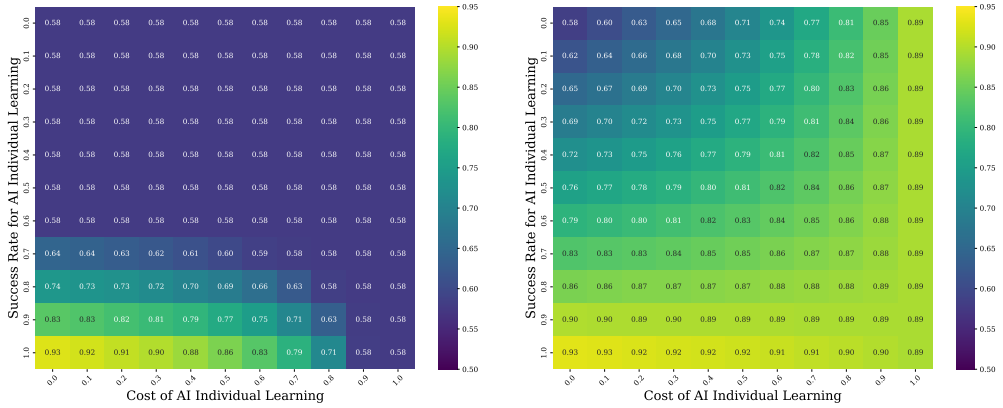

Figure 9: Allowing the AI system to either individually learn or socially learn on each turn. Impact on collective world understanding (color of cells), depending on the cost to the AI individually learning (x axis) and the expected quality of the AI individually learning about the world (y axis). Left: baseline network wherein agents either individually learn or socially learn from the AI; Right: network where agents critically social learning from the AI.

and model-centric strategies demands further study, especially as it may inform whether it is worth "paying" for more particular kinds of AI updates or modes of learning. These questions grow ever more pertinent when considering potential relative risks of "agentic" AI systems [10].

Specifically, we then extend the base simulations to modify how the AI system learns at each step. The AI system either learns by snapping to the mean of the population, or "pays" a cost of individually learning. We vary the cost of individual learning and expected accuracy of the AI system individually learning. We assess the relative gains in the base scenario (when humans either individually or socially learn based on their reliance propensity) versus a population wherein people engage in critical social learning (i.e., they can switch strategies based on the success of social learning from the AI). We use the same parameters as in the previous sections.

D Additional Details on Negative Feedback Environment Model

We extend the base simulations by adding an additional "individual learning penalty" property to each agent. The default value of this property is $\kappa = 1$ in all previous simulations (i.e., social learning from the AI has no impact on an agent's ability to individually learn); here, we consider that each time an agent learns from the AI system, we multiply their penalty parameter by a scaling factor ($\kappa_j^0 = 1, \kappa_j^{t+1} = 0.9\kappa_j^t$ for agent j who learns from the AI). An agent's individual learning success probability is then the product of their parameter value and the base individual learning success probability in the environment ($p_{ij}^{OK} = (1 - c_i)z_i\kappa_j$). All other parameters are held the same as in the base simulations.

E Looking Ahead

We close by noting several open directions that excite us about studying human-AI interaction in the context of collaborative learning in these kind of network models, which we believe are ripe for richer representations of our uncertain, dynamic world. To support further exploration of and around AI Rogers' Paradox, we make all simulation code open-source upon publication.

E.1 Handling Different "Classes" of AI Systems

Our analyses thus far have focused on a single aggregate, abstract "kind" of AI system. However, at the time of writing, there is a burgeoning offshoot of more traditional large language models that undertake more compute at inference-time returning an output to the human (e.g., o1 [34]). These models may come with higher costs for the human to learn from (e.g., longer time for a response;

higher energy costs to providing the output), yet also produce a "higher quality" response. Just as humans then often make decisions about which humans to learn from [61, 35], we may increasingly have a second order question of not only when to engage with an AI system for learning about the world — but *which* AI system to engage with. This puts an even stronger emphasis on the human being able to appraise the relative cost of their individual learning with the costs of engaging with an AI system, and being able to understand the expected utility of the different model classes.

E.2 Transparency and Information Communication

Humans frequently employ explanations to justify our recommendations and choices when learning from each other. Our analyses here do not account for (1) what potential information humans can access about AI systems, and (2) how access to that information affects their decision to engage in social learning. More broadly, AI systems may communicate their understanding of the world to humans in an attempt to modulate when and how humans integrate an AI system into their learning process. There is a large literature on the transparency of AI systems, which could potentially be repurposed for world model communication [25, 20]. Much of this work refers to obtaining and designing explanations of the behavior of an AI system in humans [19]. The goal of such transparency is often to modulate how and when humans elect to learn from an AI system [80]. This transparency information could span from the communication of uncertainty information [4] to natural language explanations of behavior [79] to representations of the systems' world model via explicit probabilistic programs [74, 75, 13, 14], or other indications of each agents' representation of the world as it relates to how that agent may communicate with another agent [66, 67]. Thus far, our network simulations have focused on a single determiner of success: whether you can "play" the right strategy or not, based on your understanding of the world. Future work could separate out these components of deciding what to do and what is known about the world, as they relate to what you know about another agent. Transparency into an AI system's understanding of the world may also be used to deter from social learning [17, 80]. In the settings we have considered here, the "use" of an AI system is resigned in favor of human judgment and individual learning based entirely on the relative success probabilities of "playing" a certain learned strategy. In our current instantiation of a form of selective use, we have assumed humans are only provided with the option to see and disregard the AI system output; transparency information instead permits the communication of *why* an output was shown or not shown (e.g., the AI system's understanding of the world is knowingly flawed, e.g., due to infrequent updates or the uncertainty is too high). Although our simulations find that such interventions do not affect population equilibrium, future work can account for alternate veils of transparency and known confounding effects for how humans update their decisions to learn upon receiving information from an AI system during learning [9, 5]. We may also consider varying degrees of information richness passed between agents about the world, which could include, for example, soft labels that capture a human's uncertainty on the task at hand [12, 65].

E.3 Collaborative Learning with Human-AI Thought Partners

The bulk of analyses related to Rogers' Paradox focus on agents as "lone explorers" engaging with the world, or agents learning hierarchically from another agent (akin to a child learning from a parent or student learning from a teacher). However, humans also learn about the world by *partnering* with each other [71] or other forms of "peer learning" [42]. Much good science has been done by two or more scientists thinking together [77] and even interaction between a teacher and student is not unidirectional [11]. Partnering often involves some kind of bidirectional *intentionality*: reasoning about the others' goals and what they know and do not know [72]. Humans and AI systems though do not necessarily have the same structure to any kind of "world model" [73, 48], ability to communicate understanding in mutually compatible ways [39, 13, 45], or costs around modes of engagement. What possibilities (and risks) may arise for different flavors of co-learning agents? One could imagine extending the negative feedback modifications we made in Section 4 to explore positive feedback, e.g., where co-learning with an AI system may boost an agent's ability to individually learn and vice versa.

E.4 Lowering Barriers to Human-Human Social Learning

Our simulations also spotlight the question around who you choose to partner with and when: if your goal is to develop a faithful understanding of the world, when do you engage in partnering over

vertical social learning versus individual learning? Presently, good human partnering is partly limited by access: in science, it can be hard, though a gift, to find good collaborators. And the human partners that we find may generally be "like" us in many ways, likely a byproduct of how we found the partner in the first place (e.g., same lab or university). The introduction of AI systems into networks of human learners may not only impact collective world model building by increasing the ease of access of good AI-based thought partners to human learners to ameliorate challenges of finding good human thought partners — AI systems in these networks can also make it *easier* to engage with a wide range of human thought partners. For instance, humans have, to an extent, been limited by how many other humans we can engage with at the same time in learning. AI systems, however, raise the prospect that humans can actually learn more efficiently from many humans. For instance, an AI system could summarize many different humans' responses, as in [69], allowing humans to consider the thoughts of many other humans more quickly. Future work can explore incorporating other actions of AI systems (e.g., as summarizers) in these agent-based population models.

E.5 Environmental Change and AI

The above possibility (introducing AI systems to change the relative costs of human interaction) underlies another important direction of study building off the Rogers' Paradox-inspired network models we consider here: what happens when an AI system changes the environment by its "presence"? Thus far, we have considered how the introduction of AI systems into networks of human and AI is impacted by changes in the environment – when we assume unidirectional impact from environment change on the world model of an agent. However, it is possible – and one may argue, even the current state of society – where changes from the AI system *change the rate of change in the environment*. Future work could explore the network effects of such dynamics.

E.6 Revisiting our Definition of "Fitness" and Notion of a Collective World Model

We close by noting an important assumption that has been baked into our discussion thus far: that there is a "single collective world model" that can be learned and that society's overall "fitness" is based on the quality of this learned world model. This is sensible in the context of our coarse environment model: there is only "one" world with an oracle strategy for agents to deploy in this world. However, human societies are immensely diverse, and it is highly simplistic to assume that AI systems and any one human easily learn a good model of the world that captures this diversity. More likely, a collective world model will disproportionately represent some experiences and cultures over others, and an "aggregate" fitness metric washes out much of this potentially heterogeneous dispersion of understanding [38, 37]. At present, our simulations have assumed that AI systems learn from the mean of the population of learners – but in practice, AI systems (often) learn from whatever data is most available, which regularly is enriched with particular biases [40, 59]. Data collection fails to represent the global majority, who may increasingly begin to interact with and learn from AI systems in the coming half-century [81]. We urge caution in interpreting these kind of simple network models, using them to fan intuition nad further thinking, rather than a rote guide. We also offer a hopeful note that one positive use of AI systems in networks of learners, as we discuss above in the context of lowering barriers to learning from other humans, is that AI systems may enable us to scale and diversify our understanding from a broader space of humans with whom we may not otherwise have engaged.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] ,

   Justification: Sections 3 and 4 describe our computational experiments. These are expanded on further in the Supplement.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] ,

   Justification: We detail Limitations throughout the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: While the bulk of our work centers around simulations, we include details on notation, where appropriate, in the Supplement.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We give experimental details in the Supplement and will release all code upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include a zip of code in the submission and all code will be made publically accessible upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Details of our experiments are provided in the main text and supplement.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We add error bars, where appropriate. We do not run statistical tests otherwise.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: All simulations were runnable on a local CPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: We ran no human experiments.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have an extended discussion in the Supplement on broader impacts and future directions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work only contains model simulations.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our work only contains our own model simulations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We develop our own model simulations. Data/code will be released upon publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not run any human experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not run any human experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We used LLMs to help with some of the coding.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.