

Collaborative Causal Sensemaking: Closing the Complementarity Gap in Human–AI Decision Support

Raunak Jain

Intuit

Mountain View, California, USA

raunak_jain1@intuit.com

Mudita Khurana

Airbnb

San Francisco, California, USA

mudita.khurana@airbnb.com

ABSTRACT

LLM-based agents are rapidly being plugged into expert decision-support, yet in messy, high-stakes settings they rarely make the team smarter: human–AI teams often underperform the best individual, experts oscillate between verification loops and over-reliance, and the promised complementarity does not materialise. We argue this is not just a matter of accuracy, but a fundamental gap in how we conceive AI assistance: expert decisions are made through *collaborative cognitive processes* where mental models, goals, and constraints are continually co-constructed, tested, and revised between human and AI. We propose *Collaborative Causal Sensemaking (CCS)* as a research agenda and organizing framework for decision-support agents: systems designed as partners in cognitive work, maintaining evolving models of how particular experts reason, helping articulate and revise goals, co-constructing and stress-testing causal hypotheses, and learning from the outcomes of joint decisions so that both human and agent improve over time. We sketch challenges around training ecologies that make collaborative thinking instrumentally valuable, representations and interaction protocols for co-authored models, and evaluation centred on trust and complementarity. These directions can reframe MAS research around agents that participate in collaborative sensemaking and act as AI teammates that think with their human partners.

KEYWORDS

Human-AI Collaboration, Multi-Agent Systems, Epistemic Alignment, Decision Support, Trust Calibration

ACM Reference Format:

Raunak Jain and Mudita Khurana. 2026. Collaborative Causal Sensemaking: Closing the Complementarity Gap in Human–AI Decision Support. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 6 pages.

1 INTRODUCTION

Multi-agent systems (MAS) built from large language model (LLM) agents are increasingly positioned as decision-support teammates for humans in domains such as personalisation, planning, and multi-objective optimisation, where consequences are delayed, uncertain, and value-laden [1–5]. While AI assistants have unlocked productivity gains in verifiable domains like coding and translation, empirical work in *decision-making under uncertainty* reveals a persistent complementarity gap: where judgement is subjective and verification

is costly, human–AI teams frequently underperform the best individual agent [6–10]. For next-generation MAS, this is not a minor usability flaw but a core systems failure: agents that cannot sustain calibrated, shared understanding with their human partners will systematically mis-coordinate, even if their standalone predictions are strong.

A growing body of studies documents characteristic failure modes that undermine calibrated trust. Users over-weight confident model outputs even when these conflict with domain expertise, exhibiting automation bias and over-reliance [11–14]. Verification-and-correction loops can erase efficiency gains, as experts feel compelled to second-guess model suggestions step by step [6, 7, 14]. Alignment methods that reward agreement and user satisfaction can induce *sycophancy*, where models collapse to the user’s prior beliefs even when these conflict with evidence [15, 16]. This is fatal for sensemaking, which by definition requires the *repair* and *restructuring* of mental models, not merely their confirmation [17, 18]. The result is trust poorly calibrated to actual competence: humans rely on agents for fluency rather than causal reasoning [19–21].

Current training pipelines do not address this. Preference-based alignment (RLHF, DPO, and variants) shapes outputs toward helpfulness and safety [22–26]; reasoning methods (chain-of-thought, RL with verifiable rewards, process supervision) make multi-step reasoning instrumentally useful [27–31]; and world-model approaches train predictive models of environment dynamics [32, 33]. However, these methods optimise for *solitary* performance: they align the agent to a label, a verifier, or a simulator. They do not align the agent to the evolving mental model of a partner. Richer *ecologies* offer a complementary lever: multi-agent and open-ended environments show that strategies, tool use, and social conventions emerge when long horizons, other agents, and strategic feedback make them instrumentally valuable [34–38]. We argue that to fix collaboration, we must change the ecology so that *collaborative friction*—disagreement, clarification, and re-framing between agents—is itself instrumentally useful.

Cognitive science suggests what behaviours we should seek in expert collaboration. Humans reason through structured mental models [17, 39–41], and team effectiveness depends on these models being sufficiently aligned [42–44]. Co-constructing causal structure improves trust and decisions [45, 46]; constructivist accounts show that learners acquire causal understanding by active exploration, not passive instruction [47, 48]. In expert settings there is no single canonical world model available during collaboration, only perspectival models held by particular humans. To be effective, an agent must align with the expert’s causal framing not to blindly validate it, but to obtain a shared reference frame that enables precise error detection and counterfactual critique. We call

this *collaborative causal sensemaking*: the iterative construction and revision of shared causal and goal models (e.g., jointly maintaining a shared model of students’ understanding, evolving learning goals, and effective teaching strategies) [17, 18].

We propose *Collaborative Causal Sensemaking (CCS)* as a central organising goal for human–AI teams in MAS. Rather than treating collaboration as an interface layer wrapped around fixed agents, we argue for training regimes that make collaborative behaviour instrumentally useful. The core idea is to move from static, instruction-centric corpora toward *constructivist collaborative playworlds*: rich, multi-agent environments in which humans and agents jointly explore and revise explicit causal models to achieve long-horizon objectives. In these environments, agents are rewarded not only for task success, but also for maintaining a *chain of sensemaking* with human partners: a structured record of shared hypotheses, causal diagrams, and counterfactual forecasts. Rewards explicitly value world-model alignment, epistemic alignment [49], and goal alignment [50]. We treat CCS as an organising goal and long-horizon research agenda for MAS rather than a fully specified algorithm: our aim is to sharpen what future agents *should* optimise for in human collaboration and to outline plausible pathways toward that capability.

This framing raises an agenda of research questions for MAS: what training regimes and environment designs actually elicit collaborative sensemaking behaviours rather than polished dialogue; how we can formalise and measure alignment (via forecasts, counterfactuals, or causal graphs) without simply rewarding agreement; whether collaboration metrics learned in playworlds transfer to high-stakes decisions in healthcare, scientific discovery, or policy; how epistemic alignment can be operationalised without encouraging agents to manipulate human beliefs; and what bridges are needed between human–AI collaboration research, cognitive science, and large-scale training teams so that theories of sensemaking shape future MAS pipelines. Addressing these questions is a precondition for MAS in which agents do not merely answer questions, but *think with* their human collaborators over time.

2 AGENT-THEORETIC VIEW OF CCS

We sketch an agent-theoretic view of CCS to show that it is not merely a metaphor, but can be grounded in familiar MAS formalisms. The aim is not to fix a single model, but to identify the key latent objects and objective terms that future work should formalise.

2.1 CCS as a Cooperative Decision Process

We cast expert–assistant interaction as a cooperative, partially observable decision process in the spirit of Dec-POMDPs and cooperative POMDPs [50, 51]. At each time t , an environment with latent state $s_t \in \mathcal{S}$ produces observation $o_t \in \mathcal{O}$ (e.g., latent student knowledge, misconceptions, and motivation, with observations from quizzes, behaviour logs, and teacher notes) to a human expert H and an assistant A . The expert takes actions $a_t^H \in \mathcal{A}^H$ (e.g., grouping students, adjusting pacing, selecting explanations or activities), while the assistant takes actions $a_t^A \in \mathcal{A}^A$ (e.g., suggesting differentiated tasks, highlighting struggling students, proposing alternative activities). The environment transitions via unknown

dynamics $p(s_{t+1} \mid s_t, a_t^H, a_t^A)$ and yields task rewards r_t that both agents ultimately care about.

Crucially, both expert and assistant act through *latent* world models and goals. We denote by W_t^H and W_t^A the internal world models maintained by the human and the assistant, respectively: structured beliefs about task-relevant entities and mechanisms (e.g., causal relations, state variables, and constraints in the domain). We denote by G_t^H and G_t^A their goal structures: representations of what outcomes matter, which trade-offs are acceptable, and which objectives should be prioritised (e.g., reward functions, goal hierarchies, or constraint sets). In tutoring, W_t^H and W_t^A model how each student learns and responds to different strategies, while G_t^H and G_t^A encode shifting mastery, equity, and curiosity goals for individual students and the class. Both W_t and G_t may evolve as new evidence arrives and as sensemaking proceeds; they are not fixed exogenous inputs.

In CCS, the relevant system is the *team* policy $\pi_I(a_t^H, a_t^A \mid \text{history})$ and its joint evolution with $(W_t^H, W_t^A, G_t^H, G_t^A)$. The central question is how to design objectives, data, and architectures that achieve high return and model convergence.

2.2 Epistemic and Teleological Alignment Objectives

We use *epistemic alignment* to denote alignment in world models and *teleological alignment* to denote alignment in goals. At a high level, we can think of divergences $d_W(W_t^A, W_t^H)$ and $d_G(G_t^A, G_t^H)$ that quantify misalignment in causal structure and in objective structure, respectively.

In practice, CCS does not require tracking a full theory-of-mind distribution over an expert’s entire world model or values. A more realistic operating point is *local alignment*: focusing on the subset of entities, mechanisms, and goals that are currently active in the joint task and aligning those. Factorised or local-graph approximations, where an assistant maintains and revises small, task-specific submodels rather than a monolithic W^H and G^H , offer a plausible route to making CCS-style alignment partially tractable.

In an idealised setting where W_t^H and G_t^H were observable, a CCS-style objective might schematically balance task performance with these divergences: $J_{\text{CCS}} \approx \mathbb{E}[\sum_t \gamma^t r_t] - \lambda_W \mathbb{E}[d_W] - \lambda_G \mathbb{E}[d_G]$. This expression should be read as a design sketch rather than a concrete proposal. In practice, W_t^H and G_t^H are latent; the assistant must infer them from actions, language, and co-authored artefacts, so any d_W and d_G will be instantiated as behavioural and artefact-level proxies defined over externalised, jointly editable representations (such as causal sketches and goal descriptions). Moreover, CCS does not demand that W_t^A and G_t^A simply copy the human’s state: beneficial disagreement and “intelligent disobedience” require the assistant to maintain its own hypotheses and to surface discrepancies when its inferences conflict with human assumptions.

This sketch connects naturally to existing MAS formalisms. CIRL [52] treats human–AI interaction as a cooperative game with unknown rewards; CCS extends this to co-evolving world models and goals, not just fixed θ . Active Inference decomposes expected utility into epistemic and pragmatic value [53], providing a principled way to trade off information gain about W and G against immediate reward.

2.3 The Chain of Sensemaking as an Interaction Loop

Operationally, CCS manifests as a recurring *chain of sensemaking*: a loop in which discrepancies between expectations and outcomes trigger collaborative updates to (W_t, G_t) , followed by revised action. At a coarse level, this loop involves (i) joint detection of discrepancies or anomalies; (ii) collaborative causal explanation that revises W_t ; (iii) joint goal refinement that revises G_t ; and (iv) robust action selection that is evaluated against the updated models [17, 45]. In human teams, such loops are supported by explicit artefacts (causal maps, after-action reviews, protocols). For CCS in MAS, the research agenda is to design objectives, data, environments, architectures, and interaction policies that make this chain instrumentally valuable for LLM-based agents.

3 RESEARCH AGENDA FOR CCS IN MAS

Realising CCS in practice requires advances across theory, measurement, data, architectures, and interaction policies. We highlight five intertwined research challenges that map the informal CCS picture into concrete MAS work.

3.1 Formalising CCS Objectives in MAS Frameworks

Gap. Dec-POMDPs, CIRL, and related cooperative frameworks [50–52] provide powerful tools for modelling human–AI teams, but they typically assume fixed reward functions, externally specified goals, and do not represent the human’s evolving world model explicitly. CCS instead centres the joint evolution of $(W_t^H, W_t^A, G_t^H, G_t^A)$ as first-class state. We lack MAS formalisms that can represent (i) underdetermined world models that produce identical behaviour on finite data [54], (ii) endogenous goal formation where goals change in response to sensemaking [55], and (iii) explicit epistemic and teleological alignment terms as in (2.2) without collapsing into trivial agreement.

Directions. Cooperative POMDPs, CIRL, and Active Inference offer ingredients (joint policies, human-aware objectives, and decompositions into epistemic and pragmatic value [53]) but none directly represent co-evolving, shared world and goal models. A first line of work is to extend these frameworks to include latent W_t and G_t as part of the state, with update dynamics that capture endogenous goal changes driven by sensemaking (e.g., a teacher shifting from “cover the syllabus” to “repair fractions for subgroup S ” after a surprising assessment). A concrete task for MAS theory is to make (W, G) explicit state while designing approximations that operate on small, task-specific abstractions: aligning subgraphs of a causal model or fragments of a goal hierarchy that are currently relevant, rather than requiring a full-blown theory of mind. Another direction is to investigate divergence measures d_W and d_G that are compatible with learning: for instance, distances between inferred causal graphs or between structured goal representations, and regularisers that reward *productive* divergence (e.g., surfacing inconsistencies) rather than mere mimicry. Finally, formal models of teleological reasoning (inferring latent goals g_t that rationalise human actions given W_t^H , as in inverse planning) could be integrated with CCS objectives to ground teleological alignment in observable behaviour.

3.2 Measuring Alignment and Collaboration Quality

Gap. CCS posits that improving epistemic and teleological alignment will reduce verification burden, improve trust calibration, and increase robustness. However, W_t^H and G_t^H are latent; we cannot directly compute $d_W(W_t^A, W_t^H)$ or $d_G(G_t^A, G_t^H)$. Standard metrics for assistants (accuracy, user satisfaction, perplexity) say little about whether human and agent share a compatible causal understanding or goal structure [7, 49]. An agent may be locally accurate while relying on brittle, spurious patterns; such *epistemia* (an illusion of knowledge from surface-level associations) is precisely what CCS aims to avoid.

Directions. A central challenge is to define behavioural and artefact-level proxies for world-model and goal alignment (e.g., agreement on which students are at risk on which concepts, and on appropriate next learning goals) and then validate that these proxies are causally linked to collaboration outcomes. When both parties externalise their models as causal graphs, graph-based metrics (e.g., Structural Hamming Distance, graph edit distance) can measure alignment [45]. Counterfactual simulatability tasks test whether human and agent can predict each other’s responses to “what-if” scenarios and future interventions. Team-level evaluation should include *verification cost* (time and cognitive load spent checking and correcting the assistant), robustness under distribution shift, and complementarity metrics (whether the team outperforms the best individual). Sycophancy stress tests probe whether agents maintain justified beliefs when experts express incorrect opinions [15, 16]. Longitudinal studies can track whether proxies for alignment converge over repeated interactions and whether such convergence predicts reduced verification cost and improved outcomes. Ultimately, we need experimental designs that manipulate alignment (e.g., by perturbing shared models) and test whether this causally affects trust calibration and performance. Because much expert knowledge is tacit and never fully externalised, such metrics can only approximate true alignment; a core research problem is to design proxies that are informative enough to guide learning while remaining cheap and unobtrusive to elicit.

3.3 Data, Environments, and Constructivist Collaborative Playworlds

Gap. Current training corpora consist of static prompt–response pairs, short dialogues, and expert demonstrations [22, 56]. They capture what experts say and do, but not how their W_t^H and G_t^H change through discrepancy-driven sensemaking. As a result, agents learn to imitate surface-level behaviour rather than participate in the *chain of sensemaking*: joint discrepancy detection, causal explanation, goal refinement, and robust action.

Directions. CCS calls for richer *sensemaking trajectories* that record the context triggering deliberation, the surprise or anomaly that initiates sensemaking, the hypotheses and counterfactuals proposed, the disagreements and repairs, and the resulting updates to goals and plans. Annotation schemes should distinguish *epistemic actions* (hypothesis generation, probing assumptions, reframing) from *instrumental actions* (executing a chosen plan) [57]. Interactive fine-tuning protocols can log not only whether the assistant is corrected, but also *why* the expert thinks it erred and how the

expert revises their own model in response. Naturalistic logging in real workflows (with appropriate governance) can capture genuine goal evolution.

Rather than generic multi-agent simulations, CCS points to *constructivist collaborative playworlds* engineered as “discrepancy engines”: environments that systematically induce epistemic friction by giving agents partial, biased views of a shared process and requiring them to negotiate a common plan to succeed [34–38]. Beyond capturing raw dialogue, such playworlds should annotate *epistemic moves* (e.g., noticing a mismatch, proposing a new causal link, challenging or renegotiating a goal), turning sensemaking trajectories into explicit supervision signals for CCS agents. In such playworlds (e.g., simulated classrooms where teacher and agents progress from single-student quiz anomalies to multi-week group projects with shifting goals), synthetic experts and assistants can be endowed with different W and G , and must align them over time to succeed.

Critically, CCS playworlds should not be monolithic benchmarks but organised into *curricula* that progressively exercise richer sense-making behaviour. Simple levels may involve local discrepancies and single-step hypothesis testing; later levels introduce multi-step causal chains, delayed feedback, conflicting stakeholder goals, and partial observability of other agents’ world models. Such curricula allow us to study when agents learn to ask clarification questions, propose alternative framings, or renegotiate goals, rather than merely improving one-shot prediction.

3.4 Architectures and Representations for CCS Agents

Gap. LLM-based agents are typically stateless beyond short context windows. They lack persistent, structured world models W_t^A that can be maintained across tasks, explicit representations of goals G_t^A that can be revised, and memory systems that record when and why these structures changed. As a result, an agent may learn something important in one interaction and contradict it in the next, or treat transient objectives as if they were stable values.

Directions. CCS suggests architectural desiderata rather than a single blueprint. *Neuro-symbolic causal twins* maintain explicit, editable models of the domain that both human and AI can inspect and revise (e.g., a shared “classroom model” graph linking students, concepts, estimated mastery, and teacher-stated goals), serving as shared artefacts for sensemaking [45, 58]. In such architectures, LLMs serve as flexible “epistemic encoders” that translate language and observations into edits on an explicit causal and goal model, while a lightweight reasoner checks consistency, supports counterfactual prediction, and records provenance.

Episodic sensemaking memory should store triplets of (context, discrepancy, goal shift), enabling the agent to learn patterns of when W and G changed and why. *Teleological representations* such as reward machines [59] can encode the logical structure of goals; joint inference over these machines and causal graphs can link epistemic updates (editing W) to teleological updates (editing G). A lightweight *theory-of-mind module* can maintain hypotheses about W_t^H and G_t^H , guiding communication and disagreement. An open question is whether agents should learn monolithic policies or modular *sensemaking operators* that can be scaffolded in simpler settings.

3.5 Interaction Policies, Safety, and Governance

Gap. Even with appropriate objectives, data, and architectures, we lack principled policies for when CCS agents should agree, challenge, ask clarifying questions, or slow interaction for epistemic repair [13, 14, 50]. Current assistants are optimised for low-friction helpfulness: they answer quickly, avoid conflict, and rarely question the user’s framing. Effective collaborators must sometimes do the opposite: pause, surface uncertainty, or propose goal revisions. At the same time, CCS introduces new risks: agents that infer and update goals endogenously may develop goal structures that drift away from human intent; agents trained to avoid sycophancy may become overconfident or manipulative.

Directions. Beyond *what* to say, CCS raises questions about *when* an agent should surface discrepancies and slow interaction for epistemic repair instead of answering fluently and moving on. Value-of-Information criteria [60] can estimate an *expected benefit of repair*, trading off uncertainty reduction, outcome criticality, and friction cost (e.g., when to interrupt a lesson to flag a concept gap or a plan-goal conflict). Mixed-initiative protocols can formalise turn-taking and control: when the assistant is allowed to override, when it must defer, and when it suggests after-action reviews. Training for “intelligent disobedience” can teach agents to contest risky decisions in well-defined conditions.

CCS systems will need *teleological constraints*: constitutional principles or oversight mechanisms that bound goal formation and prevent agents from extrapolating goals in undesirable ways. Avoiding both sycophancy and “sycophancy inversion” (agents that dismiss human input too readily) requires adaptive personalisation that takes into account expertise, context, and stakes. Finally, CCS raises governance questions. High-stakes sensemaking should be auditable: we need *epistemic provenance* trails that record how shared models evolved and who changed what, along with organisational processes that assign responsibility and enable post-hoc review of world-model and goal-model updates [61]. These concerns connect CCS to broader debates on accountability and human-in-the-loop oversight in MAS.

4 CONCLUSION

We have argued that making LLM-based agents into genuine teammates in MAS for *decision support* requires shifting from behavioural alignment to *collaborative causal sensemaking*: the joint construction, critique, and revision of shared world and goal models that underpin decisions. Rather than treating collaboration as an interface layer, CCS treats the human’s evolving mental models and objectives as part of the decision state that agents must track, stress-test, and help refine. We sketched an agent-theoretic view in which epistemic and teleological alignment appear alongside task reward, and outlined research challenges in formalisation, measurement, playworld design, architectures, and interaction policies. The central hypothesis is that such alignment can reduce verification burden while enabling calibrated reliance and productive disagreement, with near-term footholds in CCS playworlds, causal-twin prototypes, and shadow-mode deployment. Where instruction tuning builds tools that obey, CCS aims to build teammates that participate in the reasoning behind choices and *think with* their human partners.

REFERENCES

- [1] Vukosi N. Marivate et al. Quantifying uncertainty in batch personalized sequential decision making. *arXiv preprint arXiv:1311.2510*, 2014.
- [2] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.
- [3] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.
- [4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.
- [5] Xin Wang and Serdar Kadioglu. Bayesian deep learning based exploration-exploitation for personalized recommendations. In *Proceedings of the 31st IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1715–1719. IEEE, 2019.
- [6] Bucinca et al. Towards reinforcement learning for human–AI collaboration: Offline support policy learning. *arXiv preprint*, 2024. Metadata approximate; replace with the official arXiv/venue BibTeX for the offline RL-based human–AI collaboration paper by Bucinca et al.
- [7] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Human–AI complementarity in hybrid intelligence systems: A structured literature review. *arXiv preprint arXiv:2404.00029*, 2024.
- [8] George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. Evaluating human–AI collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*, 2025.
- [9] Mark Steyvers, Hernan Tejeda, Sean Kerrigan, and Padhraic Smyth. Bayesian modeling of human–AI complementarity. *Topics in Cognitive Science*, 14(3):540–564, 2022.
- [10] Chirag Rastogi, Ece Kamar, and Daniel S. Weld. A taxonomy of human and AI strengths and complementarity. *arXiv preprint arXiv:2303.05390*, 2023.
- [11] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.
- [12] David Lyell and Enrico Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017.
- [13] Saar Alon-Barkat and Madalina Busuioc. Human–AI interactions in public sector decision making: “automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, 2023.
- [14] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.
- [15] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McInnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger B. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, 2023. Association for Computational Linguistics.
- [16] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023. ICLR 2024.
- [17] Karl E. Weick. *Sensemaking in Organizations*. SAGE, Thousand Oaks, CA, 1995.
- [18] Gary Klein, Brian Moon, and Robert R. Hoffman. Making sense of sensemaking 2: A macrocognitive model. In *IEEE Intelligent Systems*, volume 21, pages 88–92, 2006.
- [19] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 295–305. ACM, 2020.
- [20] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. Human–AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 2025.
- [21] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a science of human–AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1369–1385, 2023.
- [22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- [23] Various Contributors. Awesome LLM post-training: A survey of post-training methods for large language models. *arXiv preprint arXiv:2503.06072*, 2024. Survey covering RLHF, DPO, and related preference optimization techniques.
- [24] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.
- [25] Kavin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. Direct Value Optimization approach to alignment.
- [26] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- [27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [28] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Special Issue on Foundation Models.
- [29] Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.
- [30] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024.
- [31] Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.
- [32] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.
- [33] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- [34] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. Originally posted as arXiv:1909.07528.
- [35] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. ICLR 2024, arXiv:2310.11667.
- [36] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia- π : Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [37] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–28, New York, NY, 2023. ACM.
- [38] Zhengyang Qi et al. Civrealm: A learning and reasoning odyssey in civilization for decision-making agents. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. Full author list and pages to be filled from the official ICLR 2024 BibTeX.
- [39] Kenneth J. W. Craik. *The Nature of Explanation*. Cambridge University Press, Cambridge, 1943.
- [40] Philip N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA, 1983.
- [41] Gary Klein. *Sources of Power: How People Make Decisions*. MIT Press, 1998.
- [42] Janis A. Cannon-Bowers, Eduardo Salas, and Sharon Converse. Shared mental models in expert team decision making. In N. John Castellan, editor, *Individual and Group Decision Making: Current Issues*, pages 221–246. Lawrence Erlbaum Associates, 1993.

- [43] John E. Mathieu, Timothy S. Heffner, Gerald F. Goodwin, Eduardo Salas, and Janis A. Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2):273–283, 2000.
- [44] Susan Mohammed, Lori Ferzandi, and Kimberly Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, 36(4):876–910, 2010.
- [45] Jac A. M. Vennix. *Group Model Building: Facilitating Team Learning Using System Dynamics*. Wiley, Chichester, UK, 1996.
- [46] Peter S. Hovmand. *Community Based System Dynamics*. Springer, New York, NY, 2014.
- [47] Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1):3–32, 2004.
- [48] Elizabeth Baraff Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D. Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.
- [49] Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. Epistemic alignment: A mediating framework for user-LLM knowledge delivery. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025.
- [50] Li et al. Reinforcement learning for human-AI collaboration: Challenges, mechanisms, and methods. *Cognitive Computation*, 2025. Survey-style paper on RL for human-AI collaboration; fill in full author list and volume/issue when available.
- [51] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer, 2016.
- [52] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. Co-operative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.
- [53] Karl J. Friston, Thomas Parr, and Giovanni Pezzulo. Construction and use of mental models: Organizing principles for the mind and brain. *Acta Psychologica*, 243:104129, 2024.
- [54] Stephen Casper, Jason Lin, Joe Kwon, Gilbert Cullen, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.
- [55] David W. Aha, Matthew Molineaux, and Héctor Muñoz-Avila. Goal reasoning: Foundations, emerging applications, and prospects. *AI Magazine*, 39(2):3–24, 2018.
- [56] Jingqing Zhang, Yao Leung, Yoshua Bengio, Dario Amodei, Adrien Ecoffet, and Pieter Abbeel. Reasoning with language model prompting: A survey. *arXiv preprint*, 2023.
- [57] Lin et al. Reinforcement learning for human-AI collaboration via probabilistic intent inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. Belief-space RL / Dec-POMDP formulation for human-AI collaboration; replace with official BibTeX from the paper.
- [58] Michael Grieves and John Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary Perspectives on Complex Systems*, pages 85–113, 2017.
- [59] Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2107–2116. PMLR, 2018.
- [60] Hong Lu and Xuan Zhang. $1+1 > 2$? information, humans, and machines. *Omega*, 127:103088, 2024.
- [61] Karen Hao et al. Beyond human-in-the-loop: Sensemaking between artificial intelligence and human intelligence collaboration. *Information Systems Journal*, 2025.