AI Evaluation Should Work With Humans

Jan Kulveit
Alignment of Complex Systems Group
Charles University

Gavin Leech
LCFI
University of Cambridge

Tomáš Gavenčiak
Alignment of Complex Systems Group
Charles University

Raymond Douglas
Mila

Abstract

We argue that the dominant paradigm of AI evaluation, which focuses on autonomous superhuman performance and so an implicit goal of replacing humans, is guiding AI development in the wrong direction. Instead, the AI community should pivot to evaluating the performance of human–AI teams. We argue that this collaborative shift in evaluation will foster AI systems that act as true complements to human capabilities and therefore lead to far better societal outcomes than the current process.

1 Introduction

By convention, in ML research we typically consider a problem 'solved' when the best system autonomously surpasses a human baseline (Wei et al., 2025). AI progress has largely been fuelled by this autonomous-competitive paradigm, of benchmarks that pit solo AIs against solo humans, with the implicit expectation that a successful system will work alone and replace human effort. While this Replacement Paradigm has driven innovation on many tasks, we argue that the NeurIPS community and AI research at large must reorient AI evaluation, shifting from the current focus on benchmarks that measure solo AI against solo human performance – with the tacit aim of replacing humans at each task – to a new emphasis: evaluating the *collaborative* intelligence of human–AI teams. We argue that this is necessary for systems to gain crucial prosocial capabilities (Section 2.2).

The current emphasis on human-replacement metrics, while useful for cheaply gauging certain AI capabilities, inadvertently steers development towards AI systems as substitutes for human labor and cognition (Manheim and Garrabrant, 2018). This trajectory risks exacerbating economic inequalities, devaluing human skills, and ultimately leaving humanity disempowered (Autor et al., 2015; Brynjolfsson and McAfee, 2017; Jin et al., 2025; Kulveit et al., 2025). Furthermore, it often fails to capture or promote progress on the broader capabilities essential to real-world productivity, such as asking questions, building shared understanding, and the ability to foster human creativity or insight (Kamar, 2016; Braun et al., 2023; Schmutz et al., 2024).

This paper advocates for a shift towards evaluating human-AI collaborations, where the primary question is not "*Can an AI do this task instead of a human?*" but rather "*How much more effectively can a human (or team), achieve a goal in concert with an AI?*" We will explore the limitations of the current evaluation paradigm, articulate the benefits of focusing on human-AI team performance, propose key metrics and research directions for this new focus, address potential counter-arguments, and conclude with a call to action for the community to spearhead the reorientation. By prioritizing the development and evaluation of AI that enhances human agency and collective intelligence, we can guide AI's trajectory towards more beneficial and humane futures.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

2 The Limits of the Replacement Paradigm

The prevailing paradigm in AI evaluation, characterized by its focus on beating human performance on discrete tasks, has undeniably enabled rapid advancements in algorithmic capabilities (Russakovsky et al., 2015). Benchmarks across diverse domains — from natural-language understanding (e.g., GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021)) to complex game-playing (e.g., Go (Silver et al., 2016), StarCraft (Vinyals et al., 2019)) — often use human performance as the primary yardstick, celebrating systems that operate autonomously and, ideally, achieve superhuman results. While instrumental in demonstrating AI's potential, this human-replacement focus carries significant limitations and potential perils that warrant a critical reevaluation.

2.1 Economic and Societal Misalignments

A primary concern is the socioeconomic trajectory fostered by an evaluation framework that implicitly or explicitly prioritizes AI as a substitute for human labor. Economic theory suggests that technologies can act as either complements to or substitutes for labor (Acemoglu and Restrepo, 2018, 2019). Technologies that complement human skills tend to increase human productivity, create new tasks, and can lead to higher wages and employment. Conversely, technologies that primarily substitute for human labor can displace workers, depress wages for skills that become automatable, and exacerbate income inequality (Autor et al., 2015; Brynjolfsson and McAfee, 2017).

Benchmarks and evaluation steer AI progress (Russakovsky et al., 2015). By predominantly benchmarking AI in a manner that emphasizes its capacity to perform tasks *instead of* humans, the research community inadvertently steers innovation towards substitutive applications. This can lead to an "automation race" where the primary goal becomes matching human output at lower cost, rather than exploring how AI can empower humans to achieve more or tackle entirely new challenges (Jin et al., 2025). The societal consequences of widespread labor displacement, without commensurate creation of new, human-centric roles or robust social safety nets, are a significant concern (Kulveit et al., 2025) that an evaluation paradigm focused on human replacement fails to address. Indeed, focusing on how AI can augment human capabilities may lead to more widely shared prosperity and a more optimistic vision for the future of work.

2.2 A Narrow Conception of AI Progress

The human-replacement focus, while providing clear, quantifiable metrics, can lead to a narrowed conception of what constitutes "AI progress." Success is often equated with outperforming humans on existing tasks, potentially neglecting other dimensions of intelligence, particularly those crucial for effective interaction and collaboration. For instance:

- *Interpretability and explainability for users:* While explainable AI (XAI) is a growing field, benchmarks rarely evaluate the quality or utility of explanations from the perspective of a human collaborator trying to understand, trust, or debug an AI's behavior within a shared task.
- *AI inquisitiveness and expression of uncertainty:* systems that can identify gaps in their own knowledge, ask clarifying questions, or effectively express uncertainty are crucial for robust human–AI teaming (Li et al., 2011; Zhou et al., 2019; Zhang et al., 2023; Bansal et al., 2021); yet current benchmarks mostly reward confident, definitive answers.
- *Adaptability to human intent:* Effective collaborators must adapt to partners' cognitive state, expertise, and goals, but mainstream benchmarks seldom assess this adaptability. (Consider the intuitive way in which human colleagues sense when it is 'a bad time' for a particular task, picking up on emotional cues or the timing after other events.)
- *Fostering human learning:* An AI partner could ideally help humans learn or gain deeper insights, yet few benchmarks incentivize the development of mentor-like systems.

Empirical studies corroborate these concerns: Bansal et al. show that maximizing standalone AI accuracy can paradoxically *decrease* overall human-AI team utility because humans struggle to predict or calibrate to highly complex models (Bansal et al., 2021). This narrow focus risks developing AI systems that are "brilliantly competent" in isolation but "socially inept" or difficult to integrate into complex human workflows where collaboration and shared understanding are paramount (Braun et al., 2023; Schmutz et al., 2024).

2.3 Overlooking Emergent Risks and Benefits of Human–AI Systems

A singular focus on autonomous AI performance can also lead to overlooking both emergent risks and unique benefits that arise specifically from human-AI interaction. As an example of good practice, evaluations of "dual-use" or "dangerous capabilities" (e.g., misuse for bioweapon design or misinformation (OpenAI, 2023)) are, in essence, evaluations of a human-AI team, albeit with a focus on negative outcomes. Focusing on the AI alone would greatly underestimate the risks.

Conversely, synergistic benefits—such as the creative breakthroughs observed when humans collaborate with AI in cooperative settings like Hanabi (Siu et al., 2021) or in large-scale collective intelligence contexts (Cui and Yasseri, 2024)—would be entirely missed if AI is only tested in isolation. An evaluation framework centered on human-AI teams is better positioned to identify and navigate these interaction effects.

3 What Human+AI Evaluation Could Look Like

To counteract the limitations of a purely human-replacement paradigm, we advocate for a significant shift towards the evaluation of human-AI teams, or "cyborg" systems. This approach does not seek to diminish the importance of understanding standalone AI capabilities but rather to complement them by assessing how effectively AI can work *with* humans to achieve shared goals. The central evaluative question becomes: How much more capable, efficient, insightful, or creative can a human partnered with an AI be, and what AI attributes foster such synergistic collaboration?

3.1 Defining Human+AI Team Evaluation

Human+AI team evaluation assesses the performance, interaction dynamics, and emergent capabilities of a combined system comprising one or more human users and one or more AI agents working towards a common objective. Unlike benchmarks that isolate the AI component, this paradigm explicitly considers the human as an integral part of the system being evaluated. The focus is on the *synergy* achieved: can the team accomplish what neither the human nor the AI could achieve alone, or achieve it significantly better? Recent work in cooperative game settings showcases that such evaluations are possible (Siu et al., 2021).

3.2 Key Metrics for Human+AI Teams

Evaluating human-AI teams requires a richer set of metrics beyond simple task accuracy or speed. We propose a three-pronged approach:

3.2.1 Team Task Performance

- *Outcome Quality* (accuracy, creativity, utility).
- *Efficiency* (time, resources).
- *Robustness* across conditions.
- *Innovation* or novelty of solutions.

3.2.2 Human-Centric Outcomes

- Satisfaction and Experience (e.g., SUS (Brooke, 1996), USE (Gao et al., 2018)).
- Skill Enhancement and Understanding.
- Cognitive Load Management (Sweller, 1988).
- Trust and Reliance Calibration (Lee and See, 2004; Schmutz et al., 2024).
- Sense of Agency and Empowerment.

3.2.3 Collaborative Fluency

- Interaction efficiency.
- Shared awareness and common ground (Braun et al., 2023).
- Adaptability and responsiveness.
- Error management and resilience.

3.3 Fostering Undervalued AI Capabilities

A shift towards human-AI team evaluation would naturally incentivize capabilities that current benchmarks undervalue:

- Questioning skill. How can a system prompt its human collaborator in the most fruitful way, à la active learning?
- Scaffolding human learning.
- Stimulating curiosity and exploration.
- Bidirectional explainability.

By focusing on these richer criteria, we can guide AI development towards systems that are not just powerful, but also effective, trustworthy, and empowering partners for humanity.

4 Related Work

4.1 Human–AI Teaming in Machine Learning

Early "hybrid-intelligence” systems showed that dynamically routing subtasks between algorithms and crowd workers can outperform either party alone, e.g. Kamar (2016), Lasecki et al. (2013)'s *Legion* framework for real-time crowdsourced control, or Amershi et al. (2014)'s interactive machine-teaching loop. Subsequent work generalised the pattern: in computer vision, 'double-reading' pipelines pair a CNN with a radiologist to slash miss rates while preserving throughput (Esteva et al., 2017); in games, agents explicitly optimised for cooperative play (e.g. Carroll et al. (2019)'s Hanabi-NET) achieve far higher human–AI team scores than self-play-only baselines. Parallel strands explore explanation and uncertainty communication as levers for collaboration—see Ribeiro et al. (2016)'s LIME, Senoner et al. (2024), or Bhatt et al. (2021)'s calibrated confidence displays.

Recent empirical work has begun to reveal the nuanced reality of human-AI collaboration. Chen and Chan (Chen and Chan, 2024) examined different collaboration modalities with LLMs in creative work, comparing “ghostwriter” modes (where LLMs assume the main content generation role) versus “sounding board” modes (where LLMs provide feedback on human-created content). They found that using LLMs as sounding boards helped non-experts achieve content quality closer to experts, while using LLMs as ghostwriters was detrimental to expert users due to anchoring effects. Similarly, Kumar et al. (Kumar et al., 2024) investigated the long-term effects of LLM use on human creativity, emphasizing the importance of designing AI tools as “coaches” rather than “steroids” or “sneakers” to prevent stifling human creativity even after AI assistance is removed. Their experiments showed that while LLMs boosted immediate performance, participants who worked with LLM-generated strategies showed diminished originality and diversity in subsequent unassisted work.

The challenge of evaluating human-AI collaboration has prompted new methodological frameworks. Fragiadakis et al. (Fragiadakis et al., 2024) developed a comprehensive framework proposing a structured decision tree to select relevant metrics based on distinct collaboration modes (AI-Centric, Human-Centric, and Symbiotic). Woelfle et al. (Woelfle et al., 2024) demonstrated that human-AI collaboration in evidence appraisal achieved accuracies of 89-96% for systematic review assessment, outperforming either humans or AI alone, though this required careful task design. Sharma et al. (Sharma et al., 2024) introduced a novel framework for measuring AI agency in collaborative tasks, based on social-cognitive theory. They identified key features through which agency is expressed in dialogue: Intentionality, Motivation, Self-Efficacy, and Self-Regulation, providing concrete metrics for evaluating collaborative capabilities often overlooked in traditional benchmarks. An influential manifesto for healthy collaboration with AIs is Kees Dupuis and Janus (2023).

However, challenges remain in realizing the potential of human-AI teams. Schmutz et al. (Schmutz et al., 2024) found that adding an AI teammate often reduces coordination, communication, and trust, with trust in AI tending to decline over time due to initial overestimation of capabilities. This highlights the importance of evaluating not just task performance but also team dynamics and collaborative fluency.

Despite this rich literature, the dominant framing remains performance motivated: collaboration is valued mainly insofar as it lifts headline metrics. Benchmarks still rank agents in isolation, so attributes like fostering user learning, sustaining calibrated trust, or preserving human agency are rarely optimised. Our position paper extends this body of work by arguing that *why* we pursue teaming—and how benchmarks steer research incentives—matters as much as raw task gains.

4.2 Technology, Automation, and Labor Economics

Task-based models in economics formalise two opposing forces. Automation displaces workers by letting capital perform existing tasks, while innovation that creates *new* human-advantageous tasks reinstates labour demand. Autor (Autor et al., 2015) documents how ICT hollowed out routine jobs yet boosted demand for abstract and service work. Acemoglu & Restrepo quantify the displacement/reinstatement balance and warn that innovation skewed toward substitution can suppress both wages and productivity growth (Acemoglu and Restrepo, 2018, 2019). Complementary technologies—what Brynjolfsson & Mitchell call ‘augmentation’—tend to raise productivity *and* employment, but the direction of technical change is endogenous to incentives (Brynjolfsson and Mitchell, 2017). Existing models treat that direction as an aggregate choice driven by factor prices or policy; they seldom probe the micro-level mechanisms—publication norms, leaderboards, benchmark design—that channel frontier ML research. By proposing benchmarks that measure the *synergy* and value-to-humans of AI, we provide a concrete lever to shift those incentives.

Our argument thus links the ML teaming literature’s empirical insights with macroeconomic discourse on *augmentative* versus *automating* innovation, filling a gap where the two fields rarely meet.

5 Alternative views and our rebuttals

Our proposition to shift the target of AI evaluation towards human-AI team performance faces some natural challenges:

5.1 Cost and complexity of human involvement

Clearly, one reason for the dominance of the autonomous-competitive style of evaluation is simple cost: it is far slower and more expensive to gain IRB approval, recruit, coordinate, and measure human performance relative to a static dataset benchmark. (This is especially so considering the validation set, which can be expected to be run many times in a typical iterative research process (Bouthillier and Varoquaux, 2020).) So bringing humans in could slow the rapid iteration cycles currently normal in AI research.

Rebuttal: While acknowledging the increased logistical demands, we argue for the following mitigating factors and overriding benefits:

- *Strategic investment:* The potential gains in developing AI that truly augments human capability and integrates safely into society justify the investment. Not making this investment risks developing powerful but misaligned or unusable technologies.
- *Amortizing costs with shared infrastructure:* The community can develop shared platforms, standardized interactive environments (e.g., "Human-AI Interaction Gyms"), and best practices to streamline human-AI evaluation and reduce the burden on individual research groups. The human labour for such evaluations can be crowdsourced or supplied in a commoditized way. We note the precedent of the successful crowdsourced LMArena leaderboard (Chiang et al., 2024) and that the clinical trial field has in recent decades developed a sophisticated market for fast outsourcing of participant recruitment and training, ‘contract research organizations’ (Mirowski and Van Horn, 2005). How to make this happen in ML is a valuable research direction.

- *Human surrogate models:* A promising research avenue is the development of AI-based 'human surrogate models' (Quesada et al., 2021; Kofler et al., 2022; Anthis et al., 2025). (One could make the analogy to the recent success of *in silico* screening of drug candidates in the biomedical sciences, Madden et al. (2020).) These models, trained on data from human interactions, could imitate human learning behavior, including typical errors, learning curves, and cognitive limitations, within specific task contexts. While not replacing real human evaluation entirely, they could enable more rapid and scalable iteration for many aspects of collaborative AI development, with periodic validation against actual human studies.
- *Phased and tiered evaluation:* Not all human-AI evaluations need to be large-scale. Simpler, more constrained interactive tasks can be used for initial assessments, with more complex evaluations reserved for systems showing promise.

5.2 Alternative view: Not 'Pure' AI Research

It could be argued that focusing on human-AI interaction shifts research away from 'core' AI challenges towards Human-Computer Interaction (HCI) and human-factors engineering.

Rebuttal: We contend that building AI capable of effective human collaboration is already not a pure AI problem. At the same time, it has a large ML research component.

- *New AI frontiers:* Achieving genuine human-AI synergy would benefit from progress in areas such as adaptive AI, explainable AI tailored for collaborators, representational alignment, AI alignment and novel human feedback paradigms. These are deeply technical AI research problems.
- *Intelligence in context:* True intelligence, whether artificial or natural, is often best understood and expressed through interaction.

5.3 Alternative view: Human Variability and Reproducibility

Human participants introduce variability (in skill, motivation, strategy, etc.) that can make benchmarks noisy and results hard to reproduce. Clearly this cuts against a core tenet of scientific progress (Donoho, 2024).

Rebuttal: While human variability is undeniable, established research methodologies can address this, and the challenge itself can spur innovation.

- *Robust experimental design:* Methodologies from human-subjects research (e.g., within-subject designs, appropriate sample sizes, clear reporting of participant characteristics, standardized protocols) can mitigate and account for variability.
- *AI adaptability:* The evaluation can focus on the AI's ability to adapt to different users or to improve team performance over time with a specific user, turning variability into a feature to test against.
- *Human Surrogate Models (Revisited):* Validated human surrogate models can offer a less noisy baseline for comparing certain AI capabilities, complementing direct human studies.
- *Multifaceted metrics:* Relying on diverse metrics (e.g. task performance, human-centric outcomes, collaborative fluency) can provide a more complete and robust picture than a single potentially noisy score.

A particularly important confounding variable in human+AI team evaluation is the skill level of the sampled humans at working with AI (Wang et al., 2023). Being able to measure this is thus on the critical path for our proposed collaborative evaluation, and so is a good place to start.

5.4 Alternative view: The Subjectivity of Good Collaboration

What constitutes "good" or "effective" collaboration can be subjective and harder to quantify than objective task performance like accuracy.

Rebuttal: While perfect objectivity is elusive, meaningful and actionable metrics for collaboration quality are achievable.

- *Operationalization:* We can operationalize aspects of good collaboration (e.g., efficient communication, error detection rate, shared attention) into observable behaviors.
- *Validated subjective measures:* Standardized and validated questionnaires from HCI and psychology can reliably measure subjective experiences like satisfaction, cognitive load, and trust.
- *Combining objective and subjective data:* A combination of objective task outcomes (e.g., team success, time) and subjective process measures provides a comprehensive view of collaborative efficacy.

6 Conclusion

We believe that these challenges, while real, are surmountable. They are in fact exciting research opportunities, of a similar scale to those the field has risen to in the past. Moreover, the imperative to develop AI that works *with* humans outweighs the difficulties associated with evolving our evaluation paradigm.

References

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., and Bernstein, M. (2025). Llm social simulations are a promising research method.

Autor, D. H., Dorn, D., and Hanson, G. H. (2015). Untangling trade and technology: Evidence from local labour markets. *The Economic Journal*, 125(584):621–646.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., and Xiang, A. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’21, page 401–413, New York, NY, USA. Association for Computing Machinery.

Bouthillier, X. and Varoquaux, G. (2020). Survey of machine-learning experimental methods at neurips2019 and iclr2020. Research report hal-02447823, Inria Saclay Ile de France.

Braun, M., Greve, M., and Gnewuch, U. (2023). The new dream team? a review of human-ai collaboration research from a human teamwork perspective. In *ICIS 2023 Proceedings*. Association for Information Systems.

Brooke, J. (1996). SUS: A ’Quick and Dirty’ Usability Scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation In Industry*, pages 189–194. Taylor & Francis.

Brynjolfsson, E. and McAfee, A. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. W. W. Norton & Company.

Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Chen, Z. and Chan, J. (2024). Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12):9101–9117.

Chiang, W.-L., Angelopoulos, A., Zheng, L., Sheng, Y., Dunlap, L., Chou, C., Li, T., Frick, E., Jain, N., Li, D., et al. (2024). Chatbot Arena. *LMArena, https://lmarena. ai.*

Cui, H. and Yasseri, T. (2024). AI-enhanced collective intelligence. *Patterns*, 5(11):101074.

Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*, 6(1).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Fragiadakis, G., Diou, C., Kousiouris, G., and Nikolaidou, M. (2024). Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*. Preprint.

Gao, M., Kortum, P. T., and Oswald, F. L. (2018). Psychometric evaluation of the USE (Usefulness, Satisfaction, and Ease of use) questionnaire for reliability and validity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 1414–1418.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv preprint arXiv:2009.03300 (2020).

Jin, W., Vincent, N., and Hamarneh, G. (2025). AI for just work: Constructing diverse imaginations of AI beyond "replacing humans".

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16*, pages 4068–4072. IJCAI/AAAI Press.

Kees Dupuis, N. and Janus (2023). Cyborgism. Alignment Forum.

Kofler, F., Ezhov, I., Fidon, L., Horvath, I., de la Rosa, E., LaMaster, J., Li, H., Finck, T., Shit, S., Paetzold, J., Bakas, S., Piraud, M., Kirschke, J., Vercauteren, T., Zimmer, C., Wiestler, B., and Menze, B. (2022). Deep quality estimation: Creating surrogate models for human quality ratings.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. (2025). Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv preprint arXiv:2501.16946*.

Kumar, H., Vincentius, J., Jordan, E., and Anderson, A. (2024). Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. *arXiv preprint arXiv:2410.03703*. Preprint.

Lasecki, W. S., Miller, C. D., Kushalnagar, R., and Bigham, J. P. (2013). Legion scribe: real-time captioning by the non-experts. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA. Association for Computing Machinery.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2011). The KWIK-SCIAN algorithm for active learning in supervised classification. In Gordon, G. J., Dunson, D. B., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, volume 15 of *JMLR Proceedings*, pages 449–457. JMLR.org.

Madden, J. C., Enoch, S. J., Paini, A., and Cronin, M. T. (2020). A review of in silico tools as alternatives to animal testing: principles, resources and applications. *Alternatives to Laboratory Animals*, 48(4):146–172.

Manheim, D. and Garrabrant, S. (2018). Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*.

Mirowski, P. and Van Horn, R. (2005). The contract research organization and the commercialization of scientific research. *Social studies of science*, 35(4):503–548.

OpenAI (2023). GPT-4 System Card. Technical report, OpenAI.

Quesada, C., Kostenko, A., Ho, I., Leone, C., Nochi, Z., Stouffs, A., Wittayer, M., Caspani, O., Brix Finnerup, N., Mouraux, A., et al. (2021). Human surrogate models of central sensitization: A critical review and practical guide. *European Journal of pain*, 25(7):1389–1428.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Schmutz, J. B., Outland, N., Kerstan, S., Semmer, N. K., Tschan, F., Hunziker, S., and Grote, G. (2024). AI-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*, 58:101837.

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., and Netland, T. (2024). Explainable AI improves task performance in human–ai collaboration. *Scientific Reports*, 14(1):31150.

Sharma, A., Rao, S., Brockett, C., Malhotra, A., Jojic, N., and Dolan, B. (2024). Investigating agency of llms in human-ai collaboration tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1968–1987, St. Julian's, Malta. Association for Computational Linguistics.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Siu, B., Floryan, M., Wang, H., Wu, A., Zhang, A. X., Littman, M. L., Dragan, A. D., and Critch, A. (2021). Evaluating the robustness of collaborative agents. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2021)*, pages 1227–1235. International Foundation for Autonomous Agents and Multiagent Systems.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gülçehre, c., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 3261–3275.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, B., Rau, P.-L. P., and Yuan, T. (2023). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & information technology*, 42(9):1324–1337.

Wei, K., Paskov, P., Dev, S., Byun, M. J., Reuel, A., Roberts-Gaal, X., Calcott, R., Coxon, E., and Deshpande, C. (2025). Model evaluations need rigorous and transparent human baselines. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Woelfle, T., Hirt, J., Janiaud, P., Kappos, L., Ioannidis, J. P. A., and Hemkens, L. G. (2024). Benchmarking human-ai collaboration for common evidence appraisal tools. *Journal of Clinical Epidemiology*, 175:111533.

Zhang, E., Kiseleva, J., Awadallah, A. H., Bennett, P. N., and Fourney, A. (2023). When should we ask for clarification? a study on the utility of clarification questions in conversational AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1147–1160. Association for Computational Linguistics.

Zhou, A., Choi, J., Chen, K., Tamar, A., and Abbeel, P. (2019). Uncertainty-aware proactive interaction for human-in-the-loop reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 7561–7570. PMLR.