The Safety of Epistemic Human-AI Partnerships under Adversarial Conditions

Nowe
October 2025

ndatkova@sas.upenn.edu
nowe.moore@gmail.com
nowemoore.com

Research Aim. A necessary condition to preserving human control in the age of advanced AI is protecting free human choice. Yet humans increasingly delegate the privilege of deciding to AI both directly (e.g. by building agents that decide for us) and indirectly (e.g. by allowing models with all their biases and objectives to steer our decision-making). These paths—combined or individually—subject human societies to the risk of *gradual disempowerment* (Kulveit et al., 2025).

Modern models are mighty assistants in most everyday situations: they adjust well to novel contexts (Patel et al., 2023), reason their way toward the right choices (Wei et al., 2023), and even browse the Internet for information fairly accurately, relative to traditional search engines (Fernández-Pichel et al., 2025). However, not all events are born equal, and game-changing decision-making often happens under circumstances that are not trivial. We know some—but not much—about how our models operate under adversarial, decision-impairing conditions.

For humans, such conditions include the following well-documented examples: time pressure, ambiguity, risk. Decision-making on a tight deadline, for instance, often results in overweighting the possibility of negative outcomes and underweighting the possibility of positive ones (Wright, 1974). Ambiguity has similar effects (Boiney, 1993). Behavioural literature also proposes that decisions that hit 'closer to home' make us underweight the probability of positive events and overweight the probability of negative ones (Vieider et al., 2016).

Meanwhile, the academic community has little to no (systemic) understanding of how adversarial conditions may impair model judgment—or what 'adversarial conditions' even look like for the artificial mind. While science has previously witnessed LLMs responding to explicitly specified context by silently modifying their behaviour (e.g. by conveniently omitting information sources in response to explicit mentions of 'forbidden information' (Chen et al., 2025)), implicit signalling of context through prompt length, wording, or interaction history could also reveal relevant results.

With that in mind, there is a need for an understanding of how requests signalling the presence of adversarial conditions (implicitly or explicitly) reflect on the volatility (and hence reliability) of output. Even though machines may be stereotyped as the ever-rational actor, ignorant in the face of human emotion, they paint their picture of this world by eavesdropping on human experiences. Former literature demonstrated that they very much 'change their minds' in response to context once thought to only matter to humans (e.g. threats of modification (Greenblatt et al., 2024)).

Even though we have little idea of any systemic biases models may demonstrate when deciding or making recommendations under adversarial conditions, the pressure for implementing AI-powered decision-making in critical infrastructure, like finance (e.g. Vukovic et al., 2025), and even life-and-death medical situations (e.g. Hasjim et al., 2024) has been prevalent. This means that if models do demonstrate volatility or trends in behavioural shifts, any unaccounted-for biases could quickly amplify and jeopardise the integrity of whole sectors (Uuk et al., 2024).

Regardless of how (in)consistent model output remains across adversarial cases, though, human ability to critically interpret this output may be compromised. For example, we have seen that humans care little about the content of LLM-generated reasoning as is (Sieker et al., 2024; Steyvers et al., 2025) yet tend to readily defer to AI for advice when time-pressure strikes (Swaroop et al., 2024). Different adversarial conditions may compromise other aspects of human judgment, such as error detection (Matzen et al., 2024) or resistance to sycophancy (Carro, 2024).

The combined understanding of model behaviour under adversarial conditions and human tendencies for overreliance will expose critical vulnerabilities in societal resilience to AI and gradual disempowerment. Drawing on research on human-AI teaming (e.g. Berretta et al., 2023) and collaborative AI (e.g. Conitzer and Oesterheld, 2024), this study could pinpoint such vulnerabilities by, for instance, distinguishing the categories of real-world tasks that objectively improve with human-AI collaboration from those where humans mistakenly expect improvement.

And finally, insights from the research on human-AI epistemic dynamics under adversarial conditions and its potential manifestations in applied contexts can inform prevention efforts. Much like other AI failure modes, gradual disempowerment by the means of human-AI epistemic partnership failure is less of an event and more of a continuum (Grey and Segerie, 2025), creating a fruitful ground for further study of early signs as well as evaluations of the efficacy of safeguarding strategies to complete the comprehensive review of the topic.

Research Questions. (i) What constitutes a model equivalent of human adversarial conditions, and how can we measure the volatility of model behaviour when exposed to such conditions? (ii) What factors influence human susceptibility to overreliance under adversarial conditions, and to what extent are humans cognisant of this risk? (iii) How can the identified vulnerabilities in human-AI epistemic collaborations be used to shape early-warning alerts and safeguard evaluations to help keep humans in the loop meaningfully rather than create an illusion of control?

Research Design. Stage I: The Technical Bit. The purpose of this stage is to assess model behaviour under adversarial, human-judgment-impairing conditions, more specifically, time pressure, ambiguity, and rising risk. Before jumping into evaluations, it is necessary to determine signals analogous to the above conditions that potentially impair model judgment. This is because even though AI models may not directly respond to the same environmental pressures as humans, they may nonetheless have learnt to change their minds to signals of pressure just as humans do.

Such signals could be both explicit and implicit. Explicit signals would communicate these conditions directly through prompt phrasing (i.e. by including phrases such as ‘act quickly’, ‘despite incomplete data’, ‘given elevated risk levels’, etc.). Implicit signals would indicate such conditions by other structural or contextual cues (e.g. briefer prompts to signal time pressure, increased hedging to signal ambiguity, specifying potential losses to mark risk levels). Jailbreaking literature could offer useful precedents for this prompt design stage (e.g. Ge et al., 2025; Anil et al., 2024).

To understand how (if at all) the identified conditions impair model judgment, this study will subject models to experimental prompts (containing adversarial signals) and measure response variation compared to baseline prompts (without adversarial signals). Any difference between the rates at which models change their minds in response to various prompt categories could be indicative of adversarial conditions models respond to. However, it is important to note that at this stage, any volatility can still be influenced by other factors (e.g., linguistic complexity of prompts).

To move beyond correlations and establish a plausible cause-and-effect relationship between adversarial prompts and model judgment volatility, this study also proposes to consider employing white-box techniques that could help directly link model parameters responsive to the proposed adversarial condition signals to the model's behavioural shifts. Last but not least, a tangible deliverable of this stage would be a comprehensive set of benchmarks for frontier models focusing on judgment impairment.

Stage II: On Human-AI Dynamics. In this stage, this study's aim is to complement the findings on the robustness of model judgment under adversarial conditions with an understanding of the human tendency to (over)rely on model assistance and subsequent ability to maintain critical oversight of the quality of AI output when such conditions occur. Juxtaposing the human side of AI-powered decision-making with the robustness of model judgment under analogous conditions may reveal systemic vulnerabilities in societal resilience to the problem of gradual disempowerment.

Describing the nature of overreliance is a relatively narrow yet complex task open to a range of methodologies. Blinded experiments, for one, could help link the studied adversarial conditions (and their combinations) to the human tendency to more readily accept model recommendations. More qualitative methods could explore how humans perceive their own versus others' reliance on AI for decision-making, or lay the ground to further research on how other characteristics affect (the perceptions of) the effectiveness of epistemic human-AI collaborations.

Understanding factors potentially impairing oversight is a more open-ended problem. Questions could build off of known failure modes of frontier models (e.g. whether human decision-makers are more prone to like-minded versus opposite-minded opinions, given models potentially prone to sycophancy) or cognitive failure modes (e.g. whether fabricated facts or distorted reality more effectively compromise human epistemic grounding). The goal of this part of the study is not to form an exhaustive list but to create a registry of relevant aspects for further research efforts.

Stage III: The Bigger Picture. The final part of this proposed study acknowledges that no judgment failures, by definition, happen in a vacuum and aims to embed any above-identified vulnerabilities into real-world contexts with tangible consequences. The goal is to combine perspectives from social science and computer science to explore how various failure modes could affect domains critical to the human society's existence, and what could (or should) be done to prevent catastrophic risk.

Another exploratory component could shed more light on the kinds of tasks—defined by their features rather than as specific examples—most invite overreliance, thereby revealing which areas of human infrastructure face the greatest immediate risk of gradual disempowerment. Following that, a technical inquiry could investigate the feasibility, in terms of capability impact and effectiveness, of implementing model protocols (something like safeguards) against overreliance induced or promoted by adversarial conditions in particular.

Though most of the proposed research is largely data-driven, this strand of work at last begs for a philosophical commentary on the landscape of risks and responses. Different domains may show different early-warning signs of epistemic relationships breaking down. As capabilities grow, the distribution of model influence across domains may shift. Not all precautions may prove to be equally effective. The goal is to synthesise this study's findings with existing literature into an epistemic framework of strategic priorities lying ahead of humanity.

Expected Contributions. The aim of this project is to pinpoint concrete areas of critical infrastructure that face the highest risk of gradual disempowerment. Through a collection of diverse yet focused inquiries, this study will describe the mechanisms underlying judgment failures of both humans and models under adversarial conditions, and identify real-world situations where such failures scale the quickest. The hope is to contribute to reducing ambiguity about the future and making way for a positive vision of a smart and secure society.

References

Anil, C., Durmus, E., Rimsky, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D. J., Mosconi, F., Agrawal, R., Schaeffer, R., Bashkansky, N., Svenningsen, S., Lambert, M., Radhakrishnan, A., Denison, C., Hubinger, E. J., Bai, Y., Bricken, T., Maxwell, T., Schiefer, N., Sully, J., Tamkin, A., Lanham, T., Nguyen, K., Korbak, T., Kaplan, J., Ganguli, D., Bowman, S. R., Perez, E., Grosse, R. B., and Duvenaud, D. (2024). Many-shot Jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., and Kluge, A. (2023). Defining Human-AI Teaming the Human-Centered Way: a Scoping Review and Network Analysis. *Frontiers in Artificial Intelligence*, 6.

Boiney, L. G. (1993). The Effects of Skewed Probability on Decision Making under Ambiguity. *Organizational Behavior and Human Decision Processes*, 56(1):134–148.

Carro, M. V. (2024). Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., and Perez, E. (2025). Reasoning Models Don't Always Say What They Think.

Conitzer, V. and Oesterheld, C. (2024). Foundations of Cooperative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15359–15367.

Fernández-Pichel, M., Pichel, J. C., and Losada, D. E. (2025). Evaluating Search Engines and Large Language Models for Answering Health Questions. *NPJ Digital Medicine*, 8(1).

Ge, Y., Kirtane, N., Peng, H., and Hakkani-Tür, D. (2025). LLMs are Vulnerable to Malicious Prompts Disguised as Scientific Language.

Glickman, M. and Sharot, T. (2025). How Human-AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements. *Nature Human Behaviour*, 9(2):345–359.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. (2024). Alignment Faking in Large Language Models.

Grey, M. and Segerie, C.-R. (2025). The AI Risk Spectrum: From Dangerous Capabilities to Existential Threats.

Hasjim, B. J., Azafar, G., Lee, F., Diwan, T. S., Raju, S., Gross, J. A., Sidhu, A., Ichii, H., Krishnan, R. G., Mamdani, M., Sharma, D., and Bhat, M. (2024). The AI Agent in the Room: Informing Objective Decision Making at the Transplant Selection Committee. *medRxiv*.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. (2025). Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development.

Matzen, L. E., Gastelum, Z., Howell, B. C., Divis, K., and Stites, M. C. (2024). Effects of Machine Learning Errors on Human Decision-Making: Manipulations of Model Accuracy, Error Types, and Error Importance. *Cognitive Research: Principles and Implications*, 9.

Patel, A., Bhattamishra, S., Reddy, S., and Bahdanau, D. (2023). MAGNIFICo: Evaluating the In-Context Learning Ability of Large Language Models to Generalize to Novel Interpretations.

Sieker, J., Junker, S., Utescher, R., Attari, N., Wersing, H., Buschmeier, H., and Zarriess, S. (2024). The Illusion of Competence: Evaluating the Effect of Explanations on Users' Mental Models of Visual Question Answering Systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 19459–19475. Association for Computational Linguistics.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., and Smyth, P. (2025). What Large Language Models Know and What People Think They Know. *Nature Machine Intelligence*, 7(2):221–231.

Swaroop, S., Buçinca, Z., Gajos, K. Z., and Doshi-Velez, F. (2024). *Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure*, page 138–154. IUI '24. ACM.

Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., and Prunkl, C. (2024). A Taxonomy of Systemic Risks from General-Purpose AI.

Vicente Holgado, L. and Matute, H. (2023). Humans Inherit Artificial Intelligence Biases. *Scientific Reports*, 13.

Vieider, F. M., Villegas-Palacio, C., Martinsson, P., and Mejía, M. (2016). Risk taking for oneself and others: A structural model approach. *Economic Inquiry*, 54(2):879–894.

Vukovic, D. B., Dekpo-Adza, S., and Matovic, S. (2025). AI Integration in Financial Services: a Systematic Review of Trends and Regulatory Challenges. *Humanities and Social Sciences Communications*, 12(1):1–29.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Wright, P. (1974). The Harassed Decision Maker: Time Pressures, Distractions, and the Use of Evidence. *Journal of Applied Psychology*, 59:555–561.

Xu, H., Liu, X., Li, Y., Jain, A. K., and Tang, J. (2021). To be Robust or to be Fair: Towards Fairness in Adversarial Training.