AI systems are rapidly scaling up and becoming both capable and cost-effective. If humans' ability to oversee them does not scale accordingly, the rush in development could result in systems that lack reliable alignment or exhibit behaviors that cannot be efficiently understood and controlled. My research aims to address this by **making human supervision of AI systems both reliable for difficult tasks and efficient for complex outputs**, ultimately ensuring their safety even as they surpass human capabilities and operate at massive scale.

PAST WORK

My journey toward this research direction began in my first year at the University of Hong Kong, where I joined Prof. Chao Huang's lab to work on **self-supervised learning and recommender systems**. I focused on improving their robustness to noise and sparsity, aiming to accurately model users with limited interaction histories. While mask-reconstruction training improved learning from noisy data in other fields, I found it ineffective when applied to our setting due to the lack of semantics in individual nodes of the user interaction graph. To address this, I proposed a path-based reconstruction objective that adaptively samples informative graph paths based on their contribution to the recommendation goal. The resulting algorithm improved recommendation accuracy, especially for users with a short interaction history, leading to my **first-authored SIGIR 2023 publication** [6]. This experience not only taught me how to identify and address limitations in existing approaches but also showed me how one's research can translate to meaningful societal impact.

Driven by a deeper interest in understanding intelligence, I then joined Prof. Yixin Zhu's Cognitive Reasoning lab at Peking University to study **computational cognitive science**. I worked on understanding how humans intuitively make predictions of the physical world. We designed novel human experiments to show that pure simulation-based theories did not fully account for humans' systematic biases. This observation led us to propose a new computational framework that conceptualized intuitive physics as a dual process involving both heuristics-based reasoning and probabilistic simulations. Beyond contributing to the experiment design and human studies, I played a key role in developing the theory through extensive discussions with the team. This ongoing work strengthened my ability to design human studies and translate empirical observations into theoretical innovations, which was essential for my later research on human oversight of AI.

A pivotal transition in my research trajectory happened when I came across the AI alignment forum. Its posts inspired me to think deeply about how future AI systems would interact with human society and why potential risks must be addressed proactively. To enter the field of AI safety, I identified a concrete challenge from the literature: the **tension between helpfulness and harmlessness in language model alignment** often leads to overly conservative models that reject even benign requests. I proposed training language models to explicitly reason about context-specific safety considerations before making decisions and reached out to Prof. Huaxiu Yao at UNC Chapel Hill for supervision on this project. My hypothesis was that as models become more capable at general reasoning, they might also become better at analyzing nuanced situations and distinguishing genuine threats from benign scenarios that resemble harmful content. While this project did not result in a publication, it taught me invaluable lessons about choosing research directions with long-term impact rather than immediate but superficial solutions.

CURRENT AND FUTURE WORK

Determined to work with leading researchers in AI safety, I visited Berkeley in Spring 2024 to join

Prof. Jacob Steinhardt's group. There, I initiated a project about **scalable oversight**: as language models become more capable, the tasks they are given become harder to supervise; will current post-training methods remain effective under **unreliable supervision**? To test this, I simulated unreliable task demonstrations and comparison feedback using small models and time-constrained humans, and surprisingly found that the canonical reinforcement learning from human feedback (RLHF) pipeline breaks down in this setting. To address this, I proposed iterative label refinement (ILR) as an alternative method. Instead of updating the *model* as in RLHF, ILR updates the *data* by using the same type of comparison feedback to decide whether human-labeled data should be replaced by model-generated proposals, then retrains the model on the updated data. By doing so, ILR allows large model updates that improve performance without suffering from overoptimization under unreliable supervision like RLHF. This work [5], **currently under review at ICLR 2025** (scored 8/8/8/5), provides key insights for aligning AI systems under unreliable supervision and enhances my ability to conduct large-scale empirical research—from systematic experiment design to rigorous analysis that yields novel understanding. Moreover, it sharpened my research taste and deepened my understanding of fundamental challenges in scalable oversight.

During my PhD and after, I aim to make human supervision of AI systems reliable on hard tasks and efficient for complex outputs. I identify three directions that together work toward this goal:

Robust alignment under unreliable supervision: I will continue my work on improving the robustness of AI alignment under realistically unreliable human supervision. As humans inevitably make mistakes due to limited capabilities or cognitive and social biases, we need methods that ensure reliable alignment even with systematically flawed supervision.

Scaling human oversight to interacting AI systems: I plan to develop algorithms and tools that enable humans to oversee widely deployed AI systems, where many instances of them interact with each other in ways that produce complex outputs with emergent risks. One solution is training simpler but specialized AI agents to help humans monitor and understand these outputs—such agents present a way to leverage massive computational power for scaling up human oversight.

Scaling laws for human oversight: I hope to conduct extensive human studies and derive scaling laws that can predict human effort required for reliable oversight of an AI system on specific tasks; this will reveal whether current methods can keep pace with rapidly advancing AI. Moreover, I aim to build on these human study results to develop AI-based simulations that approximate human scaling laws, providing a cost-efficient environment for fast testing and iteration of new oversight approaches.

WHY BERKELEY

Many faculty at Berkeley align with my research agenda. I am particularly excited to continue working with Prof. Jacob Steinhardt, whose vision of using AI to help humans understand AI [4] aligns well with my goal. I am also drawn to Prof. Anca Dragan's work on modeling human irrationality [3, 1], which is critical for robust alignment under systematically biased human supervision. Prof. Stuart Russell's research on understanding how private AI information leads to deception [2] is also invaluable for my goal of assisting humans to understand complex AI systems, because these systems may produce outputs that are too complex to be fully observed by humans (*e.g.*, OpenAI's o1 generates hidden chain-of-thought). Additionally, the vibrant Bay Area AI safety community provides an irreplaceable environment for my research and my long-term goal of becoming a professor who works to make AI safe.

REFERENCES

[1] CHAN, L., CRITCH, A., AND DRAGAN, A. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956* (2021).

[2] LANG, L., FOOTE, D., RUSSELL, S., DRAGAN, A., JENNER, E., AND EMMONS, S. When your ais deceive you: Challenges of partial observability in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.17747* (2024).

[3] SINGHAL, S., LAIDLAW, C., AND DRAGAN, A. Scalable oversight by accounting for unreliable feedback. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment* (2024).

[4] STEINHARDT, J., AND SCHWETTMANN, S. Introducing transluce — a letter from the founders, October 2024. Accessed: 2024-12-06.

[5] YE, Y., LAIDLAW, C., AND STEINHARDT, J. Iterative label refinement matters more than preference optimization under weak supervision. In *Submitted to The Thirteenth International Conference on Learning Representations* (2024). under review.

[6] YE, Y., XIA, L., AND HUANG, C. Graph masked autoencoder for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023), pp. 321–330.