

Prelude: The Vision. 💡 In E. M. Forster's 1909 short story "The Machine Stops" [14], all of humanity lives in isolated underground cells, completely dependent on an all-providing mechanical system called "The Machine", which handles all physical and social needs. "Year by year it [The Machine] was served with increased efficiency and decreased intelligence. The better a man knew his own duties upon it, the less he understood the duties of his neighbor." Many of our current AI tools bear concerning resemblance to The Machine: they make our writing faster but more predictable [3], our collective thinking less diverse [10, 2], and may unwittingly influence our opinions [18]. But *we need to think for ourselves*, for the sake of the many intellectual activities we find so important — as thoughtful moral actors, as engaged political citizens, as value-conscious scientists, as philosophers, as historians, as readers of literature and appreciators of art, as *thinking and feeling humans*. To complement today's machines that make us more *efficient*, we need "thinking tools" which help us think more *intelligently*, critically, reflectively, and that *connect us with our neighbor-humans* rather than isolating us within The Machine.

Broadly, I want to build AI/IA tools which help us do better critical thinking, 😊 particularly in *the textual social sciences* (e.g., philosophy, history, literature) and in *the public sphere*. I envision AI tools that *textual social sciences* scholars engage with to think more deeply about the human condition, and even to innovate new concepts for problems both old and new. I envision *public encounters* with AI that generate the right kind of intellectual friction to make people more thoughtful about political, cultural, and moral issues in our lives. Contributing to HCI, NLP, and social science venues, I will **perform ethnographic work** to understand how AI can support different intellectual communities, **develop modeling techniques and design human-AI interactions** for critical thinking, and **build open-access tools** for those communities. As a student in computer science and philosophy, working with computer science professors Amy Zhang and Ranjay Krishna and philosophy professor Rose Novick, **I have published several first-author papers related to these topics**, and am currently leading several related projects. I believe that the intellectual, technical, and personal skills I have developed through this work equips me to pursue this research agenda.

Note: Throughout, I use the **A** marker to flag my projects and the **I** marker to flag proposed directions.


Building intellectual friction into AI tools. 🌀 Facing the right kind of intellectual friction can make us reflect on our intentions and beliefs in productive ways [12, 5, 9]. How can we build this friction into AI models and tools?

First, **AI models must be able to represent a wide range of perspectives**, so that they can present and advocate for perspectives that users may not have encountered. I have contributed to work addressing multiple different instances of this problem. **A** To understand the different ways that AI models can be pluralistically aligned to diverse values, **I helped formulate three technical definitions of pluralism** — Overton, steerable, and distributional — and provide philosophical support, outlined in an ICML position paper [33]. The vocabulary and framework introduced in our paper has gained traction in alignment and agent modeling communities [22, 13, 4]. **B** But models should also be able to represent types of perspectives subtler than values. Vision-language model (VLM) training is biased towards English web-scale image-text pairs, with the majority of samples in popular datasets in English [26, 16, 30]. However, psychological and linguistic studies show that speakers of different languages describe images in systematically different ways [19, 24, 28]. Inspired by these studies, **I discovered distributional differences in both the content and expression of captions produced in different languages**, both in datasets and by models. I also showed that models fine-tuned on captions produced in a particular language internalize the content and expressive biases of that language. Our work [41], submitted to CVPR, highlights that VLMs should be trained on captions produced from a variety of languages to maximize exposure to different perspectives of seeing and describing the world, inspiring further work [26]. **C** In a similar vein, I spent a summer at Deepgram, a speech-to-text API, designing and optimizing data curricula — orderings of data samples based on their information characteristics (e.g., content, speed, audience). By exposing models to the right kinds of diverse data at the right time in training, I trained large speech-to-text models to achieve more robust performance with half the training time. Through these three projects, I've **gained experience developing techniques both to measure and build representation of diverse perspectives** in models.

Once models can represent a wide range of perspectives, **how should they be used to build AI tools with intellectual friction to help humans do critical thinking?** **D** To answer this question, I interviewed 21 philosophers — people who definitely do a lot of critical thinking — on their interactions with LLMs. I found that philosophers overwhelmingly had negative experiences with using LLMs to do philosophy, for two main reasons: LLMs lack *selfhood* (a persistent sense of self, bold and consistent articulation of opinions and beliefs) and *initiative* (ability to formulate intellectual interests/goals and pursue them, start conversation, introduce ideas). From these interviews, I formulated a design framework to articulate where current LLMs fall short and how future designs could succeed in becoming better critical thinking tools. **I presented this paper at COLM [42], where I had many insightful conversations**, discussing further directions in this space both for modeling techniques and interaction design. This paper has usefully framed much of the future work I want to do: both *selfhood* and *initiative* usefully describe ways of building intellectual friction into AI models and tools. **E** **I am currently leading a project to introduce greater initiative into image generation interfaces.** Although we generally want models to produce diverse images of people, the "woke Gemini" controversy [29] showed that "diversity can be taken too far". Perhaps the solution is *factuality* — images of people should be diverse only when it is factually valid (e.g., doctors, CEOs; not founding fathers, Nazi soldiers). However, *factuality is not enough*: the visual representation of people is entangled in broader cultural, political, and historical debates (e.g., what does "a German citizen" [1], "Jesus" [11, 25, 32], "a woman" [38, 7] look like?). These debates

cannot be, or at least have not yet been, settled into matters of fact. By building image generation tools with the *initiative* to engage users in these debates, image generation can become a site of reflection, rather than a blind tool to actualize our (possibly flawed or ignorant) intentions. We are currently running user studies and will submit to FAccT 2025.

In the future, I want to continue working on both the *selfhood* and *initiative* dimensions of building critical thinking tools. **1** On the *selfhood* front, **I want to build models that can convincingly represent, develop, and advocate for “uncommon sense”** — ideas that are unintuitive and contradictory to “common sense”. For example, it is common sense that “pain is bad, and we should avoid it”, whereas it may be uncommon sense that “we should sometimes put ourselves in pain to cultivate character”. Arguably, working/thinking through the latter is an important part of being human [17, 36, 21, 34]; our AI tools should support this reflective process. A large body of existing work in NLP and CV aligns models towards common sense reasoning in the contexts of physical reasoning [6], vision [43], norms/morality [44], etc. We can adapt many of these alignment techniques to develop uncommon sense in models. One conceptually simple method is to expose models to high-quality materials that we know contain rich worked examples of uncommon sense, such as in the textual social sciences. Moreover, existing alignment pipelines can be adapted to reward dispreferred (rather than preferred, as usual) outputs, inducing more provocative, possibly thought-provoking model behavior [8]. **2** On the “initiative” front, **I want to build models that can ask us piercing, stimulating questions**. This inverts the traditional human-LLM interaction where the human asks the question first, which excludes rich contexts where the human “doesn’t know what they don’t know”. One such project as an “inverse Delphi” [20]: rather than training models to *give moral judgments* on human questions (e.g., “this action is inappropriate with 95% certainty”), they adaptively *ask provocative and reflective questions* about the human’s moral problem (e.g., “have you thought about how X might be affected by this action?”). I believe this approach emphasizes human agency in moral choice-making, addressing many concerns about building “moral AI” [15, 35]. As before, I believe there is abundant (albeit overlooked) data for this task in the textual social sciences and even popular sources (e.g., NYT’s *The Ethicist*).

Concept discovery.  New concepts help us frame what we already know in a new light, but also guide how we seek out new knowledge. **F** In one of my earlier projects, I demonstrated how shifting concepts — the organizing frame through which we view a problem — can solve problems which arise with an existing concept, *in the medical imaging domain*. In the typical “singular” concept of medical segmentation annotation, a single boundary is produced around the region of interest. Existing uncertainty representation methods are difficult to interpret because they rely on singular annotations. I proposed a shift from a “singular” to a “bounding” concept of annotations, in which annotators explicitly mark high- and low-certainty bounds on the range of plausible segmentations. I found that “bounding” annotations represent the variance in singular annotations while reducing annotator disagreement, and are *preferred by practicing clinicians* over alternative representations of uncertainty based on singular annotations. **I presented this paper at AAAI HCOMP and received an honorable mention award** [40].

While we can propose new concepts to help us solve modeling problems, *we can also ask models to help us produce new concepts*. [31] proposes to “bridg[e] the human-AI knowledge gap” by using interpretability tools to introduce new concepts from humans to models. While [31] introduces new chess concepts, **I am interested in introducing new moral, cultural, and philosophical concepts with AI models**. **G** In a NeurIPS workshop paper [27], we argue that unaligned LLMs already grasp the conceptual social structure of morality because they are trained to reproduce the discourse by which moral beliefs are constructed, and therefore can teach us a lot about moral thinking. Thus, complementing existing work *training* models to learn moral values [33, 20, 44], I want to develop methods to **probe models as repositories of moral knowledge**, both old and new. **H** In a paper presented at the Pacific University Philosophy Conference [39], I further argue that the material process of engaging with vision-language models can direct our cognitive labor towards self-reflection in ethically significant ways. Because (vision-)language models learn and represent knowledge so differently from us — struggling with tasks we find basic while excelling on ones we find difficult — teaching them concepts can yield fruitful reflection on our own part.

Inspired by recent work in concept induction [23] and novel concept representation [37], I want to build **interaction-first concept discovery**. By their nature, it would be difficult to grasp moral, cultural, and philosophical concepts by passively observing the outputs of interpretability probes; rather, humans will need to actively engage with the model via the concept discovery apparatus. **3** In particular, I am excited about **conceptual vocabulary generation** for conceptual discovery. In the textual social sciences, concepts are often developed by introducing new words and developing their meaning by producing a web of texts that each cultivates some aspect of the concept. As an example, consider a new made-up word LOREM, and three sentences using it: { “*Sometimes, a stranger’s glance carries the unexpected weight of LOREM.*” , “*Politicians fear that LOREM may lead to public sympathy with extremists.*”, “*I can’t explain it — I just felt a sense of LOREM with him, like I understood him.*” } After reading this, you may infer that LOREM means something like an intuitive and spontaneous sense of human connection, but by the nature of language, the meaning of LOREM will exceed than the sum of what is written using it in these texts. As more texts using the word “LOREM” are produced, our idea of LOREM’s meaning will take on more depth and complexity. If an LLM system could successfully produce such texts using mystery words in meaningful ways, it could be a valuable tool for textual social scientists to innovate new concepts to understand the human condition.

References

- [1] Oya S. Abali. *German Public Opinion and Immigration: Tensions and Paradoxes*. Migration Policy Institute, Accessed November 25, 2024. URL: <https://www.migrationpolicy.org/sites/default/files/publications/TCM-GermanPublicOpinion.pdf>.
- [2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. "Homogenization Effects of Large Language Models on Human Creative Ideation". In: *Proceedings of the 16th Conference on Creativity & Cognition*. C&C '24. Chicago, IL, USA: Association for Computing Machinery, 2024, pp. 413–425. ISBN: 9798400704857. DOI: 10.1145/3635636.3656204. URL: <https://doi.org/10.1145/3635636.3656204>.
- [3] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. "Predictive text encourages predictable writing". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 128–138. ISBN: 9781450371186. DOI: 10.1145/3377325.3377523. URL: <https://doi.org/10.1145/3377325.3377523>.
- [4] Joshua Ashkinaze et al. *Plurals: A System for Guiding LLMs Via Simulated Social Ensembles*. 2024. arXiv: 2409.17213 [cs.CL]. URL: <https://arxiv.org/abs/2409.17213>.
- [5] Steve Benford et al. "Uncomfortable interactions". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 2005–2014. ISBN: 9781450310154. DOI: 10.1145/2207676.2208347. URL: <https://doi.org/10.1145/2207676.2208347>.
- [6] Yonatan Bisk et al. "PIQA: Reasoning about Physical Commonsense in Natural Language". In: *AAAI Conference on Artificial Intelligence*. 2019. URL: <https://arxiv.org/pdf/1911.11641>.
- [7] Judith Butler. *Gender Trouble: Feminism and the Subversion of Identity*. 1st. New York: Routledge, 1990.
- [8] Alice Cai, Ian Arawjo, and Elena L. Glassman. *Antagonistic AI*. 2024. arXiv: 2402.07350 [cs.AI]. URL: <https://arxiv.org/abs/2402.07350>.
- [9] Carl DiSalvo. *Adversarial Design*. The MIT Press, 2012. URL: <http://www.jstor.org/stable/j.ctt5hhbs4>.
- [10] Anil R. Doshi and Oliver P. Hauser. "Generative AI enhances individual creativity but reduces the collective diversity of novel content". In: *Science Advances* 10.28 (2024), eadn5290. DOI: 10.1126/sciadv.adn5290. eprint: <https://www.science.org/doi/pdf/10.1126/sciadv.adn5290>. URL: <https://www.science.org/doi/abs/10.1126/sciadv.adn5290>.
- [11] Bart D. Ehrman. *From Jesus to Christ: The Origins of the New Testament Images of Christ*. New Haven, CT: Yale University Press, 2000. URL: <https://yalebooks.yale.edu/book/9780300084573/from-jesus-to-christ/>.
- [12] Robert H. Ennis. "A taxonomy of critical thinking dispositions and abilities". In: *Teaching thinking skills: Theory and practice*. Ed. by Joan Boykoff Baron and Robert J. Sternberg. W H Freeman/Times Books/Henry Holt & Co., 1987, pp. 9–26.
- [13] Shangbin Feng et al. *Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration*. 2024. arXiv: 2406.15951 [cs.CL]. URL: <https://arxiv.org/abs/2406.15951>.
- [14] E. M. Forster. *The Machine Stops*. Originally published in the Oxford and Cambridge Review. Various later reprints and anthologies, 1909. URL: https://en.wikisource.org/wiki/The_Machine_Stops.
- [15] Timnit Gebru and Émile P. Torres. "The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence". In: *First Monday* 29.4 (Apr. 2024). DOI: 10.5210/fm.v29i4.13636. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/13636>.
- [16] Rachel Hong et al. *Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp*. 2024. arXiv: 2405.08209 [cs.CY]. URL: <https://arxiv.org/abs/2405.08209>.
- [17] Aldous Huxley. *Brave New World*. United Kingdom: Chatto & Windus, 1932, p. 311.
- [18] Maurice Jakesch et al. "Co-Writing with Opinionated Language Models Affects Users' Views". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. DOI: 10.1145/3544548.3581196. URL: <https://doi.org/10.1145/3544548.3581196>.
- [19] Li-Jun Ji, Zhiyong Zhang, and Richard E Nisbett. "Is it culture or is it language? Examination of language effects in cross-cultural research on categorization." In: *Journal of personality and social psychology* 87.1 (2004), p. 57.
- [20] Liwei Jiang et al. *Can Machines Learn Morality? The Delphi Experiment*. 2022. arXiv: 2110.07574 [cs.CL]. URL: <https://arxiv.org/abs/2110.07574>.
- [21] Søren Kierkegaard. *Fear and Trembling*. Danish. C. A. Reitzel, Oct. 1843.
- [22] Thom Lake, Eunsol Choi, and Greg Durrett. *From Distributional to Overton Pluralism: Investigating Large Language Model Alignment*. 2024. arXiv: 2406.17692 [cs.CL]. URL: <https://arxiv.org/abs/2406.17692>.
- [23] Michelle S. Lam et al. "Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLoM". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642830. URL: <https://doi.org/10.1145/3613904.3642830>.
- [24] Takahiko Masuda and Richard E. Nisbett. "Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans." In: *Journal of personality and social psychology* 81 5 (2001), pp. 922–34.

- [25] David Morgan. *The Real and the Sacred: Picturing Jesus in the Gospel of John*. Ann Arbor, MI: University of Michigan Press, 2023. URL: <https://press.umich.edu/Books/T/The-Real-and-the-Sacred>.
- [26] Thao Nguyen et al. "Multilingual Diversity Improves Vision-Language Representations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2024. URL: <https://arxiv.org/abs/2405.16915>.
- [27] Mark Pock, Andre Ye, and Jared Moore. "LLMs grasp morality in concept". In: *Proceedings of the NeurIPS 2023 Moral Philosophy and Psychology (MP2) Workshop*. 2023. arXiv: 2311.02294 [cs.CL]. URL: <https://arxiv.org/abs/2311.02294>.
- [28] Jakob Prange and Nathan Schneider. "Draw mir a sheep: A supersense-based analysis of german case and adposition semantics". In: *KI-Künstliche Intelligenz* 35.3-4 (2021), pp. 291–306.
- [29] Adi Robertson. *Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis*. The Verge, February 21, 2024. Feb. 2024. URL: <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>.
- [30] Christoph Schuhmann et al. "LAION-5B: An open large-scale dataset for training next generation image-text models". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022. URL: <https://openreview.net/forum?id=M3Y74vmsMcY>.
- [31] Lisa Schut et al. *Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero*. 2023. arXiv: 2310.16410 [cs.AI]. URL: <https://arxiv.org/abs/2310.16410>.
- [32] Jeffrey Siker. "Historicizing a Racialized Jesus: Case Studies in the "Black Christ," the "Mestizo Christ," and White Critique". In: *Biblical Interpretation: A Journal of Contemporary Approaches* 15.1 (Jan. 2007), pp. 26–53. URL: https://brill.com/view/journals/bi/15/1/article-p26_1.xml.
- [33] Taylor Sorensen et al. "A Roadmap to Pluralistic Alignment". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vienna, Austria, 2024. URL: <https://arxiv.org/abs/2402.05070>.
- [34] The Struts. *Could Have Been Me*. Music video directed by Jonas Åkerlund. Aug. 2015. URL: https://www.youtube.com/watch?v=ARhk9K_mvIE.
- [35] Zeerak Talat et al. *A Word on Machine Ethics: A Response to Jiang et al. (2021)*. 2021. arXiv: 2111.04158 [cs.CL]. URL: <https://arxiv.org/abs/2111.04158>.
- [36] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Incerto. Random House (US), Penguin Books (UK), 2012, p. 519. ISBN: 1-400-06782-0.
- [37] Ryan Teehan, Brenden Lake, and Mengye Ren. "CoLLeGe: Concept Embedding Generation for Large Language Models". In: *Proceedings of the 2024 Conference on Language Models (COLM)*. 2024. arXiv: 2403.15362 [cs.CL]. URL: <https://arxiv.org/abs/2403.15362>.
- [38] Lisa Walker. *Looking Like What You Are: Sexual Style, Race, and Lesbian Identity*. Accessed November 25, 2024. New York, NY: NYU Press, 2001. ISBN: 9780814793725. URL: <https://nyupress.org/9780814793725/looking-like-what-you-are/>.
- [39] Andre Ye. "And Then the Hammer Broke: Reflections on Machine Ethics From Feminist Philosophy of Science". In: *Pacific University Philosophy Conference*. 2024.
- [40] Andre Ye, Quan Ze Chen, and Amy Zhang. "Confidence Contours: Uncertainty-Aware Annotation for Medical Semantic Segmentation". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 11. 1. 2023, pp. 186–197. DOI: 10.1609/hcomp.v11i1.27559. URL: <https://doi.org/10.1609/hcomp.v11i1.27559>.
- [41] Andre Ye et al. *Computer Vision Datasets and Models Exhibit Cultural and Linguistic Diversity in Perception*. 2024. arXiv: 2310.14356 [cs.CV]. URL: <https://arxiv.org/abs/2310.14356>.
- [42] Andre Ye et al. "Language Models as Critical Thinking Tools: A Case Study of Philosophers". In: *Proceedings of the Conference on Language Modeling (COLM 2024)*. 2024. arXiv: 2404.04516 [cs.HC]. URL: <https://arxiv.org/abs/2404.04516>.
- [43] Rowan Zellers et al. "From Recognition to Cognition: Visual Commonsense Reasoning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6720–6731. URL: https://openaccess.thecvf.com/content_CVPR_2019/papers/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.pdf.
- [44] Caleb Ziems et al. "NormBank: A Knowledge Bank of Situational Social Norms". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 7756–7776. DOI: 10.18653/v1/2023.acl-long.429. URL: <https://aclanthology.org/2023.acl-long.429>.