Scaling Up AI Alignment

Aarti Singh
Machine Learning Department, Carnegie Mellon University
aarti@cs.cmu.edu

Abstract

From expert AI systems of the 1970s to self-supervised systems of the 2020s, the pendulum of AI development has swung from heavy reliance on human feedback to no or minimal reliance in the last 50 years. Self-supervised approaches have contributed significantly to the success and scalable development of AI. However, today we are at a tipping point where the future of AI, and whether society ends up benefiting from this technology in the long run, depends critically on the subsequent AI development aligning with human goals and values. Realizing this, there has been ramping up of efforts to align AI models with human expectations and values. Human feedback, however, remains limited and difficult to elicit. Thus, a key question lingers – how can we scale up alignment of AI systems with individual expectations and societal norms? This paper provides an overview and perspective on efforts at answering this question.

Introduction

Over the last 50 years, we have witnessed a paradigm shift in AI development with respect to its reliance on human input - from expert systems of the 1970s ([Felgenbaum 1977) when experts were heavily engaged in the design of AI systems to craft rules that were hard-coded into a knowledge base that was then queried with an inference engine, to the purely self-supervised AI systems of the 2020s (Meta 2021) where machine learning systems are trained by being asked to predict held out data (parts of an image or subsequent words in text). While the latter approach has multiplied amount of data available for training machine learning models by several orders of magnitude and enabled scalable training of AI systems on large amounts of data, the resulting systems are often found to be disconnected from human goals and values. For example, deployment of unsupervised AI systems has led to biases and unfair treatment (Reuters 2018, Haider 2024), financial losses (CNN 2021, TechHQ 2024), loss of privacy (HRD 2025), harmful (The Register 2022, APNews 2020) and unethical (APNews 2022) decisions, ultimately leading to AI adoption failure. Thus, we are at a tipping point where the future of AI, and whether society ends up benefiting from this technology in the long run, depends critically on the subsequent AI development aligning with societal goals and values.

Consequently, in the last few years, there is increasing effort in the research community to develop human-centered AI. In this paper, we take a broad view on alignment which we define as *the design of AI models, agents, and systems that contribute to the attainment of human goals and values, both at individual and societal level*. This includes both perspectives where AI system outputs should adhere to and directly optimize for individual expectations and societal norms, but also where AI systems should complement and/or assist a human to improve the value or performance on a goal set by the human individual or society.

We begin the paper by summarizing some of the key approaches proposed in recent literature for achieving alignment of AI models with human goals and values. Then, we focus on a key shortcoming – the need of human feedback for AI alignment, which is limited. Finally, we investigate the question of how to scale up alignment of AI models?

Recent Approaches to AI Alignment

While attempts at incorporating human feedback in AI models have existed for several years, AI alignment has been investigated recently mostly in the context of Generative AI systems, and particularly focusing on Large Language Models. Earlier attempts beyond supervised and active learning using human labels, included preference-based learning and optimization (Xu 2017, Xu 2020, Yue 2012, Wirth 2017), feature feedback (Raghavan 2006, Attenberg 2010), imitation learning or behavioral cloning (Hussein 2018), etc. that incorporated human feedback or behavior while training AI models, and which are often still the method of choice in

embodied robotic systems. However, almost all current attempts at aligning Generative AI systems focus on post-training alignment, mainly due to the efficacy of self-supervised systems that are trained on vast amounts of unlabeled data in a scalable manner. Key attempts in this direction include:

1. **Guardrails** – are hand-crafted rules that are deployed to filter an AI model's inputs and outputs (Dong 2024b). This is perhaps the simplest approach which can be implemented via prompting in AI systems such as Large Language Models. However, this requires significant, domain-specific expert input and guardrails are often violated, a phenomena known as jailbreaking (Deng 2024).
2. **Red teaming** – refers to the practice of setting up a team of ethical hackers to evaluate an AI model in a controlled setting by attempting to elicit misaligned information from it (Feffer 2024). This also requires significant effort and understanding of AI failure modes to craft intelligent hacks that can reveal the weaknesses of an AI system.
3. **Retrieval augmented generation (RAG)** – refers to requiring an AI model to retrieve evidence from a vetted corpus of data, typically text, at inference time to support its response (Lewis 2020). This requires expert knowledge to identify data sources that are considered trustworthy to include in the RAG knowledge base of vetted information.
4. **Alignment using human preference datasets** – is done by using an explicit or implicit reward model gleaned from fixed datasets that capture human feedback. This includes RLHF (Reinforcement Learning from Human Feedback) which explicitly learns a reward model using pairwise comparisons from humans and then fine-tunes the pre-trained policy using it (Stiennon 2020, Ouyang 2022), as well as approaches such as Direct Preference Optimization (DPO) (Rafailov 2023), IPO (Gheshlaghi 2024), GRPO (Shao 2024), etc. that implicitly use a reparametrized reward to directly update the policy. These methods also have extensions to online preference tuning to account for dynamic preferences (Dong 2024), multi-objective and multi-group settings to account for the diversity of preferences from multiple stakeholder groups (Xiong 2025), as well as multi-agent setting where AI collaborates with a human that is providing preferences on the desired behavior of an AI teammate (Vanshika 2024).

The Scalability Question

All these approaches to AI alignment described above critically depend on human feedback - in the case of safeguards, human feedback is in the form of hand-crafted rules, red-teaming requires carefully designed hacks, RAG pipelines require vetting information, and RLHF, DPO etc. require datasets of preference comparisons and ratings. All these forms of human feedback are limited and not very scalable. Thus, there is a dire need to consider alternate ways to scale up human feedback in a reliable manner.

We believe answering the scalability question requires confluence of ideas from diverse disciplines particularly social & decision sciences and AI to identify 1) what form of human feedback is easy to elicit and scale up?, and 2) can we use computational models as proxies of humans? These questions have been explored over decades of social & decision science research, and this is a call for AI researchers to leverage those insights towards finding a scalable solution to AI alignment.

Potential Directions

In this section, we sketch some directions that are either starting to be explored or can be explored to scale up AI alignment by exploring the two questions of nature of human feedback that is more amenable to scaling up, and the use of computational proxies.

First, consider the question of what form of human feedback is easy to obtain? The use of comparisons over ratings or direct evaluations in existing alignment work was inspired by the fact that it is sometimes easier for humans to compare options rather than choose one (Thurstone 1927). However, it still requires a lot of direct concerted effort. Unsolicited human feedback in the form of demonstrations, instructions or discussions of good and bad behavior is another form of feedback that has been explored somewhat for use in training AI agent, such as via behavior cloning and imitation learning (Hussein 2018), but is less explored for AI alignment and complementing humans. Another approach that is being explored is to elicit rubrics as another form of human feedback and holds great promise due to its high scalability for evaluation and alignment. While eliciting rubrics is not easy, it can improve over scalability of alignment approaches that try to explicitly or implicitly *learn* (Stiennon 2020, Ouyang 2022, Rafailov 2023, Gheshlaghi 2024, Shao 2024) rather than *elicit* a rubric/reward model. Success in eliciting rubrics also paves the way for use of surrogate evaluators such as AI models and LLM judges (Zheng 2023) that can be used to auto-grade and guide the alignment of AI models.

This brings us to the second question of computational proxies. Since learning an AI model to mimic humans tends to be expensive, there are increasing number of attempts to use Generative models and LLMs as surrogates for humans by prompting them to adopt a persona at test time (Tseng 2024, Park 2023). While an LLM persona or LLM guided by human rubrics may be able to act as a human proxy, more research needs to be done to validate if such AI models can indeed mimic human preferences and decision making (Bavaresco 2025). There are also more direct approaches to modeling human decision making via cognitive models based on decades of social & decision science research that can be considered. Some of these computational models such as cognitive architectures ACT-R (Polk 2002), iBL (Gonzalez 2023) etc. are often developed in a data-free manner or require only a few datapoints to calibrate some model parameters. Such a computational model of human decision making can act as a scalable source of synthetic, and often even dynamic, human preferences that can be queried in an on-demand fashion. On the social sciences side, extension of these cognitive models to group, population, or societal behavior rather than individual preferences and rewards would be informative to align AI with societal norms. The use of such cognitive computational models of human decisions for AI alignment has not been explored and can hold much potential.

We end this paper with a call to AI researchers and social & decision scientists to work closely together to identify scalable solutions for human alignment of AI models. It is by leveraging this interdisciplinary expertise that we can ensure subsequent AI development aligns with societal goals and values, and that the society ends up benefiting from AI technology in the long run.

Acknowledgements

This work is supported by the AI Research Institutes Program funded by the National Science Foundation under the AI Institute for Societal Decision Making (NSF AI-SDM), Award No. 2229881.

References

APNews. 2020. Backup driver in fatal Arizona Uber autonomous crash charged. https://apnews.com/article/technology-business-arizona-phoenix-homicide-fdd1574ac6a3c418d4f2b569b797dc16

APNews. 2022. How AI-powered tech landed man in jail with scant evidence. https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220

Attenberg, J.; Melville, P.; and Provost, F. 2010. A unified approach to active dual supervision for labeling features and examples. In Machine Learning and Knowledge Discovery in Databases, pages 40–55. Springer.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; Martins, A.; Mondorf, P.; Neplenbroek, V.; Pezzelle, S.; Plank, B.; Schlangen, D.; Suglia, A.; Surikuchi, A. K.; Takmaz, E.; and Testoni, A. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 238–255, Vienna, Austria. Association for Computational Linguistics.

CNN. 2021. Zillow's home-buying debacle shows how hard it is to use AI to value real estate. https://www.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html

Deng, G.; Liu, Y.; Li, Y.; Wang, K.; Zhang, Y.; Li, Z.; Wang, H.; Zhang, T.; and Liu, Y. 2024. MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots. Proceedings Network and Distributed System Security Symposium.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. RLHF Workflow: From Reward Modeling to Online RLHF A Comprehensive Practical Alignment Recipe of Iterative Preference Learning. Transactions on Machine Learning Research.

Dong, Y.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Ruan, W.; and Huang, X. 2024. Position: building guardrails for large language models requires systematic design. In Proceedings of the 41st International Conference on Machine Learning, 235. JMLR.org, Article 451, 11375–11394.

Felgenbaum, E.A. 1977. The art of artificial intelligence: themes and case studies of knowledge engineering. In Proceedings of the 5th international joint conference on Artificial intelligence - Volume 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1014–1029.

Feffer, M.; Sinha, A.; Deng, W. H.; Lipton, Z. C.; and Heidari, H. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society. AAAI Press, 421–437.

Gheshlaghi A., M.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In International Conference on Artificial Intelligence and Statistics, pages 4447–4455. PMLR.

Gonzalez, C.; Lerch, J.F.; and Lebiere, C. 2023. Instance-based learning in dynamic decision making, Cognitive Science, Volume 27, Issue 4, Pages 591-635, ISSN 0364-0213, https://doi.org/10.1016/S0364-0213(03)00031-4.

Haider, S.A.; Borna, S.; Gomez-Cabello, C.A.; Pressman, S.M.; Haider, C.R.; and Forte, A.J. 2024. The Algorithmic Divide: A Systematic Review on AI-Driven Racial Disparities in Healthcare. J Racial Ethn Health Disparities. 2024 Dec 18. doi: 10.1007/s40615-024-02237-0.

HRD. 2025. Nearly half of organisations entering employee info into GenAI: survey. https://www.hcamag.com/us/specialization/hr-technology/nearly-half-of-organisations-entering-employee-info-into-genai-survey/530858

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2018. Imitation Learning: A Survey of Learning Methods. ACM Computing Surveys 50, 2, Article 21, March.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 793, 9459–9474.

Meta. 2021. Self-supervised learning: The dark matter of intelligence. https:/ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; and Ray, A., et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

Polk, T.; and Seifert, C. 2002. Cognitive Modeling. Cambridge, Massachusetts: MIT Press. ISBN 0-262-66116-0.

Raghavan, H.; Madani, O.; and Jones, R. 2006. Active learning with feedback on features and instances. Journal of Machine Learning Research, 7.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Reuters. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.-M.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. ArXiv abs/2402.03300.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33:3008–3021.

TechHQ. 2024. Oh, Air Canada! Airline pays out after AI accident. https://techhq.com/news/air-canada-refund-for-customer-who-used-chatbot/

Thurstone, L.L. 1927. A law of comparative judgement. Psychological Review, 34, 273-286.

Vanshika, V.; et al. 2024. A survey on human-ai teaming with large pre-trained models." arXiv preprint arXiv:2403.04931.

Wirth, C.; Akrour, R.; Neumann, G.; and Fürnkranz, J. 2017. A survey of preference-based reinforcement learning methods. The Journal of Machine Learning Research, 18(1): 4945–4990.

Xu, Y.; Muthakana, H.; Balakrishnan, S.; Singh, A.; and Dubrawski, A. 2020. Nonparametric Regression with Comparisons: Escaping the Curse of Dimensionality with Ordinal Information. Journal of Machine Learning Research, JMLR, 21(162): 1-54.

Xu, Y.; Zhang, H.; Miller, K.; Singh, A.; and Dubrawski, A. 2017. Noise-Tolerant Interactive Learning Using Pairwise Comparisons. Advances in Neural Information Processing Systems.

Xiong, N.; and Singh, A. 2025. Projection Optimization: A General Framework for Multi-Objective and Multi-Group RLHF. International Conference on Machine Learning.

Yue, Y.; Broder, J.; Kleinberg, R.; and Joachims, T. 2012. The K-armed dueling bandits problem. Journal of Computer and System Sciences, 78(5).

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In Proceedings of the 37th International Conference on Neural Information Processing System.