Scopes of Alignment

Kush R. Varshney, Zahra Ashktorab, Djallel Bouneffouf, Matthew Riemer, and Justin D. Weisz
IBM Research
1101 Kitchawan Road
Yorktown Heights, NY 10598

Abstract

Much of the research focus on AI alignment seeks to align large language models and other foundation models to the context-less and generic values of helpfulness, harmlessness, and honesty. Frontier model providers also strive to align their models with these values. In this paper, we motivate why we need to move beyond such a limited conception and propose three dimensions for doing so. The first scope of alignment is *competence*: knowledge, skills, or behaviors the model must possess to be useful for its intended purpose. The second scope of alignment is *transience*: either semantic or episodic depending on the context of use. The third scope of alignment is *audience*: either mass, public, small-group, or dyadic. At the end of the paper, we use the proposed framework to position some technologies and workflows that go beyond prevailing notions of alignment.

1 Introduction

A new paradigm of artificial intelligence (AI) is upon us. Large language models (LLMs) or foundation models, to refer to the same concept irrespective of data modality more generally, are finding uses across application domains (Bommasani et al. 2021). Such use is often to generate new outputs—such as text, images, code, or molecules—based on the training data and a prompt. Although there is excitement for LLMs, and practical applications are seeing a return on investment, we must be aware of their risks and harms (Weidinger et al. 2022; Shelby et al. 2023). Traditional risks of AI, like lack of fairness, robustness, explainability, transparency, and uncertainty quantification are present in LLMs, but LLMs amplify the existing harms and introduce new harms. Some new harms include hallucination or lack of factuality, hateful speech, prompt injection, information leakage, copyright infringement, bullying and gaslighting. Thus, a major current endeavor in responsible AI beyond making LLMs as helpful as possible is to make them as harmless and honest as possible (Sun et al. 2024). LLMs are being described in terms of their behaviors using the language of psychology. Although we must be careful with respect to anthropomorphism (Shneiderman and Muller 2023), behavior is emerging as the lingua franca for specifying desired input–output relationships of LLMs and for evaluating their performance.

Values are fundamental beliefs that guide behaviors. They indicate the importance of various things and actions to a person or group of people, and determine the best ways to behave. As shown in Fig. 1, there are many ways throughout the AI application development lifecycle to get an LLM to behave in a desired fashion. Data scraping, curation and pretraining an entire model from scratch offer an approach that is often prohibitively costly in terms of both data and computation. Safeguarding, through moderations or constrained decoding based on classifiers or detectors for various harm dimensions (Achintalwar et al. 2024a), is lightweight in data and computing resources but may not entirely capture all sorts of desired behaviors.

The steps in the middle of the lifecycle between LLM pre-training and safeguarding are known these days as *alignment* (Shen et al. 2023). Alignment steps are meant to embed behaviors and values in the AI system and may be done over more than one round. (In the figure, the dotted line implies that alignment is an optional step.) However, alignment is an 'empty signifier': something without an underlying meaning (Kirk et al. 2023). Different parties have appropriated the term to refer to various goals and actions. Lipton quips (Lipton 2024): "Alignment is now defined so broadly that all of AI, all of ML, and the entire history of technology is—and always has been—'alignment research'." Gilbert quips (Gilbert 2024): "Alignment—the sensible kind anyway—is just human-centered computing." Frontier model providers typically refer to alignment as endowing LLMs with instruction-following behavior and helpful/harmless/honest values (Bai et al. 2022; Ouyang et al. 2022).

There are several ways of carrying out the goal of alignment, each having a different cost profile. Full fine-tuning changes all of the parameters of the LLM and can be done through supervised learning and/or reinforcement learning. Parameter efficient fine-tuning changes a small number of parameters, often through low-complexity adapters. Prompt engineering does not change the LLM at all, but relies on prompts to achieve the desired behavior. Additionally, system prompts — special prompts appended before all inputs — are commonly developed and sometimes made public by

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

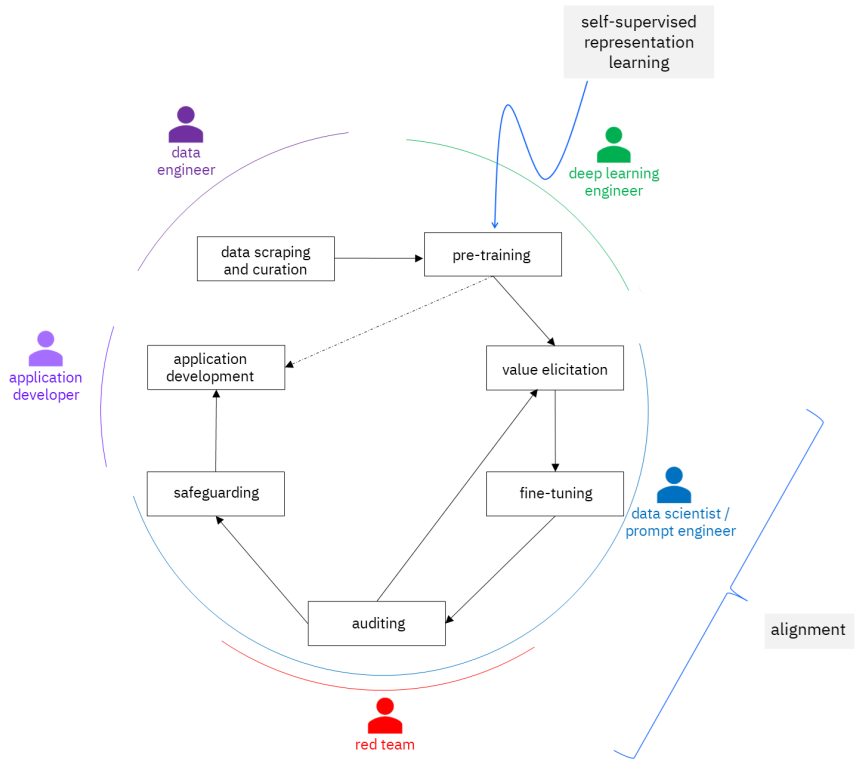

Figure 1: Typical application development lifecycle for LLMs considered by frontier model providers.

frontier model providers (Anthropic 2024).

2 Limitations of Alignment

There are limitations to the LLM lifecycle and how LLM alignment has been conceived and carried out to date. The alignment methods and approaches of frontier model providers lead to a singular set of values that are intended across all contexts, all deployments, and all time. Many recent empirical studies show the set of knowledge and values within LLMs is sociopolitically biased toward dominant cultural sources (Durmus et al. 2023; Feng et al. 2023). The root cause is alignment, as much as, or more than pre-training (Qi et al. 2023).

Moral psychology is the key sub-area within psychology that deals with values. Morality is the differentiation of behaviors between those that are right (proper) and those that are wrong (improper); it can be a body of principles derived from a particular religion or culture. It can be highly personal as well. Morality captured by multilingual language models does not reflect cultural differences, but is dominated by high-resource languages and cultures (Haemmerl et al. 2023; Scherrer et al. 2024). Such analysis has been organized around moral foundations theory, which sets up six dichotomies: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression (Graham et al. 2013).

Clear patterns show up empirically through the lens of moral foundations theory, but we may ask if there is something more basic going on. The prevailing approach to alignment seems to espouse the silver rule — do not do to others what you would not like them to do to you — and the golden rule — do to others what you would like them to do to you. The focus is completely on the person doing the action (agent) and their perception, not on the person to whom the action is done (patient). In contrast, the platinum rule that foregrounds the patient and their perception — do to others what *they* would like done to them — is currently not adequately incorporated in alignment. Yet, all three are important.

Dyadic morality theory, an alternative to moral foundations theory, specifically considers moral issues as situations where an intentional agent is perceived to cause damage to a vulnerable patient (Schein and Gray 2018). And, it is the differences in the perception of who or what is an intentional agent, who or what is a vulnerable patient, and what constitutes a causal relationship, that lead to differences in morality. Are weather phenomena intentional agents? How about AI systems? Are unborn children vulnerable patients? Can the saying of a mantra be the cause of damage? Moreover, these perceptions may be different and change based on the type of damage, the context, or other factors. Thus, dyadic morality theory is less universal and more contextually-scoped than the moral foundations theory that undergirds prevailing views of alignment.

It is under this light of different perceptions of intentional agents causing damage to vulnerable patients that we

state the need for different scopes of alignment, i.e. different strokes for different folks. We emphasize that this call for different scopes of alignment is a precursor to dealing with *conflicting* values, skills and knowledge that is more often considered in the nascent field of pluralistic alignment (Dognin et al. 2025). We believe it should be dealt with first before going to the harder problem of simultaneously conflicting values. In fact, by scoping the alignment properly, we may even avoid the conflicts that arise in larger scopes.

3 Different Scopes of Alignment

Alignment should not be just one activity, technical approach, or workflow. It should be more specific and scoped for the different needs of different groups over time. Toward this end, we propose the following three scopes of alignment: (1) competence, (2) transience, and (3) audience.

By the *competence* of alignment, we mean which kind of capability the LLM is being given further instruction or fine-tuning in: knowledge, skills, or behaviors. Knowledge is a topically-organized set of facts and information that supports work-related performance. Skills — what the model can do — are about the model's proficiency and ability to perform a job-related activity that contributes to effective performance. Some examples of skills displayed by LLMs include summarizing emails and converting texts into hip-hop raps. Behaviors are the values, attitudes, and temperament evidenced through the actions of LLMs. Behaviors can include verbosity, politeness, succinctness, objectivity, formality, inclusiveness, respecting social norms, non-egregiousness, faithfulness, and non-profanity. In human terms, the United States Army, as part of its Army Talent Attribute Framework, counts many different knowledge, skills, and behaviors (KSBs) as essential attributes that "allow individuals to achieve goals of greater complexity and scale than they can achieve on their own" (Homer and Fleming 2023). Presumably they would also be essential attributes for AI systems to team with people to achieve their goals.

By the *transience* of alignment, we mean the boundedness to time or context. Using terminology stemming from neuroscience, we posit two categories of alignment transiences: *episodic* and *semantic*. Episodic alignment is for particular knowledge, skills or behaviors bound to a time, place, or other context. Semantic alignment is for general KSBs about the world that applies to all times and contexts. Semantic knowledge includes concepts like an understanding of the law of gravitation, whereas episodic knowledge may include the state of things in a dynamic environment, such as who is in the building right now. Semantic behaviors may include general duties that are valued across contexts, such as not stealing. In contrast, episodic behaviors may include particular behaviors from an organization's code of conduct, such as obtaining manager pre-approval before booking air tickets, or ones that are contextual based on relational ethics, such as an attorney not divulging incriminating conversations with their client. The behavior transience dichotomy of LLMs is described using the terminology sādhāraṇa-dharma and viśeṣa-dharma by (Varshney 2024). Skills may be similarly dichotomized.

By the *audience* of alignment, we mean what kind of group the alignment, and ultimately the deployed model, is for. Taking a cue from communication theory, groups can range from a single person interacting with an AI system like an AI personal assistant or therapist (dyadic), to a small group, to a community (public), to everyone (mass). As illustrated in Fig. 2, dyadic alignment between one person and one AI system is a two-way street: people and the LLM can both adapt to achieve better alignment. In recent human-computer interaction literature, this idea has been referred to as mutual theory of mind (Wang et al. 2024; Wang and Goel 2022). This can also be the case in a small-group environment with a handful of people and one or more LLM-based agents. However, public alignment and mass alignment are generally focused on the model being aligned to the needs of a community or all of humanity. The medium that distinguishes public alignment from mass alignment is analogous to the difference between simply speaking to a community in-person and requiring media such as newspapers, radio, television, or social networks to broadcast a message. In the case of mass LLM alignment, the communication is in the opposite direction and the medium needs to somehow bring in the valued knowledge, skills, and behavior from all individuals, groups, communities, cultures, and publics. It must be a representation of the full diversity of human culture without missing anything. We do not know of any such effective medium.

4 Motivating Example

A compelling example for demonstrating the need for different scopes of alignment is a (hypothetical) AI system designed to support mental health counseling in the United States and China, where cultural norms and regulatory environments differ significantly.

Competence: In the United States, where openness and individualized therapy are emphasized, AI alignment would focus on competencies such as empathic communication, offering personalized strategies, and handling a wide array of mental health issues with flexibility. In China, however, AI competence may need to align with collectivist values, prioritizing solutions that emphasize family and community-oriented approaches, as well as deference to social harmony and privacy regulations.

Transience: In the U.S., episodic alignment might be prioritized, allowing the AI to adapt dynamically to diverse individual needs and new mental health practices as they emerge. In China, semantic alignment may be favored, as the AI system aligns more consistently with established practices and cultural norms around mental health, which are more stable and slower to change in response to new trends.

Audience: The audience of alignment would also vary. In the U.S., the AI system might be tailored for individuals in private settings, including direct interactions with therapists or mental health apps. In China, the audience of alignment could target both individuals and local community health groups, with the AI's output possibly needing to be accessible to family members or other community stakeholders in keeping with collectivist support structures.

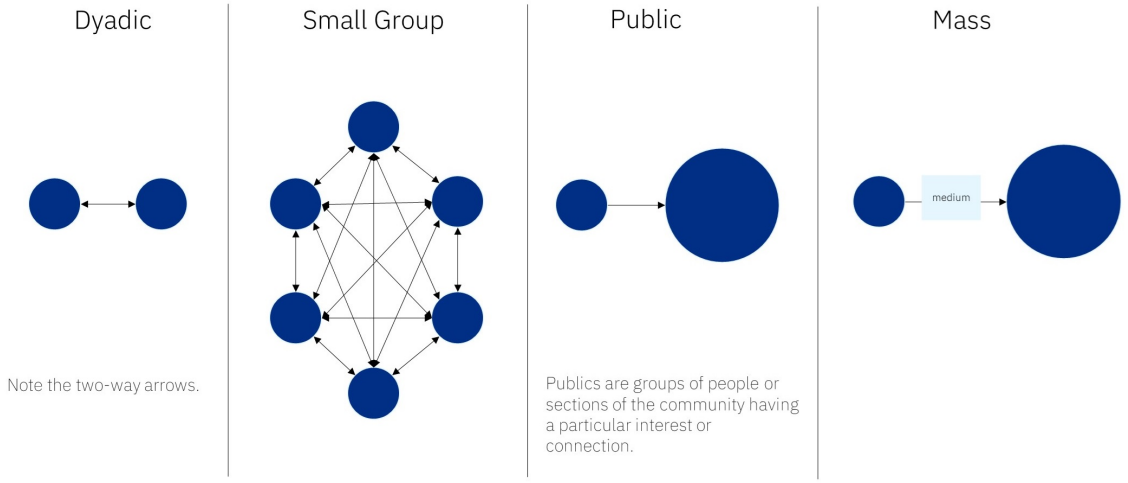

Figure 2: Different audiences of alignment.

This example illustrates why a flexible, multi-scope approach is essential for AI alignment, as country-specific social and cultural norms shape how AI should interact, evolve, and serve its intended audience.

5 Implications

What is the use of defining these three scopes of alignment? The first use is simply to be able to more precisely discern that alignment is not a singular process; it contains many possibilities. Most of the literature focuses on mass semantic alignment for behaviors (Ji et al. 2023), but there are so many other ways to conduct alignment.

The more important use is to clearly specify the technical needs of alignment technologies and workflows. As Tan Zhi-Xuan et al. submit, "Reward-based alignment—and preference matching more generally—is only appropriate for AI systems with sufficiently local uses and scopes. In other words, it is adequate for only the narrow or minimalist versions of the value alignment problem, where the values and norms at stake can be summarized as a reward function specific to the system's scope" (Zhi-Xuan et al. 2024).

For example, different competences of alignment may lead to different kinds of data requirements and optimization algorithms. Knowledge is often best learned by generating a lot of questions and answers about content rather than simply feeding in the plain content (Sudalairaj et al. 2024). However, for skills, many examples and repetition is more effective. Behaviors are better learned as preferences supplemented with scenarios in which those behaviors occur (Padhi et al. 2024). It is important to differentiate the transience of alignment because while semantic alignment implies (full) fine-tuning a model in advance and leaving it that way, episodic alignment needs to be done at inference time based on the context. Episodic alignment may be possible with adapters learned using parameter efficient fine-tuning methods that can be applied on the fly. The audience of alignment determines whether the procedure is bidirectional or unidirectional, as well as the source and format of the KSBs.

The discussion in this paper does not detail choosing the right scope of alignment, but such a process is clearly necessary. It may be a human-centered activity that aims for a reflective equilibrium in which the scope has a competence, transience, and audience with fairly small internal conflict of values (Möller 2016; Zhi-Xuan et al. 2024). We will find an appropriate scope through such an approach; we will also implicitly deal with the problem of conflicting values because the reflective equilibrium will have resolved conflicts by narrowing the scope until there are few remaining.

6 Related Work

Alignment beyond mass semantic alignment for behaviors is not a white space, and there are plenty of theory and methods that exemplify other scopes. We recount a non-exhaustive list for illustration purposes only. Along the audience dimension, personalized alignment and mutual theory of mind take us toward dyadic and small group alignment (Kirk et al. 2024; Wang et al. 2024). Alignment from unstructured text takes us toward public alignment (Padhi et al. 2024; Achintalwar et al. 2024b). Along the transience dimension, we have already discussed how low-rank adapters can contribute to episodic behaviors (Hu et al. 2022). For knowledge, there are emerging LLM architectures with episodic memory (Das et al. 2024). Along the competency dimension, we are seeing the emergence of some alignment methodologies focused on knowledge and skills (Li et al. 2024; Sudalairaj et al. 2024).

7 Conclusion

In this paper, we submit a broader conception of LLM alignment that is organized around three scopes: competence, transience, and audience. We do so not only to point out that the literature is mostly focused on mass semantic alignment for behaviors, but also to delineate some of the technical

methodologies that are required to pursue other scopes. We believe these methodologies to be a prerequisite for further work in pluralistic alignment; steerable pluralism can mediate between different scopes (Sorensen et al. 2024), but they need to have been defined and implemented first. Scopes of the right level may not even need mediation if they contain minimal internal conflict or are at a level in which value conflicts have already been deliberated upon.

References

Achintalwar, S.; Alvarado Garcia, A.; Anaby-Tavor, A.; Baldini, I.; Berger, S. E.; Bhattacharjee, B.; Bouneffouf, D.; Chaudhury, S.; Chen, P.-Y.; Chiazor, L.; Daly, E. M.; de Paula, R. A.; Dognin, P.; Farchi, E.; Ghosh, S.; Hind, M.; Horesh, R.; Kour, G.; Lee, J. Y.; Miehling, E.; Murugesan, K.; Nagireddy, M.; Padhi, I.; Piorkowski, D.; Rawat, A.; Raz, O.; Sattigeri, P.; Strobelt, H.; Swaminathan, S.; Tillmann, C.; Trivedi, A.; Varshney, K. R.; Wei, D.; Witherspooon, S.; and Zalmanovici, M. 2024a. Detectors for safe and reliable LLMs: Implementations, uses, and limitations. arXiv:2403.06009.

Achintalwar, S.; Baldini, I.; Bouneffouf, D.; Byamugisha, J.; Chang, M.; Dognin, P.; Farchi, E.; Makondo, N.; Mojsilović, A.; Nagireddy, M.; Natesan Ramamurthy, K.; Padhi, I.; Raz, O.; Rios, J.; Sattigeri, P.; Singh, M.; Thwala, S.; Uceda-Sosa, R. A.; and Varshney, K. R. 2024b. Alignment studio: Aligning large language models to particular contextual regulations. *IEEE Internet Computing* 28(5):28–36.

Anthropic. 2024. System prompts. https://docs.anthropic.com/en/release-notes/system-prompts#july-12th-2024.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; El Showk, S.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Li, F.-F.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the opportunities and risks of foundation models. arXiv:2108.07258.

Das, P.; Chaudhury, S.; Nelson, E.; Melnyk, I.; Swaminathan, S.; Dai, S.; Lozano, A.; Kollias, G.; Chenthamarakshan, V.; Dan, S.; and Chen, P.-Y. 2024. Larimar: Large language models with episodic memory control. In *Proceedings of the International Conference on Machine Learning*.

Dognin, P.; Rios, J.; Luss, R.; Sattigeri, P.; Liu, M.; Padhi, I.; Riemer, M. D.; Nagireddy, M.; Varshney, K. R.; and Bouneffouf, D. 2025. Contextual value alignment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Durmus, E.; Nyugen, K.; Liao, T. I.; Schiefer, N.; Askell, A.; Bakhtin, A.; Chen, C.; Hatfield-Dodds, Z.; Hernandez, D.; Joseph, N.; Lovitt, L.; McCandlish, S.; Sikder, O.; Tamkin, A.; Thamkul, J.; Kaplan, J.; Clark, J.; and Ganguli, D. 2023. Towards measuring the representation of subjective global opinions in language models. arXiv:2306.16388.

Feng, S.; Park, C. Y.; Liu, Y.; and Tsvetkov, Y. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. arXiv:2305.08283.

Gilbert, E. 2024. HCC is all you need: Alignment—the sensible kind anyway—is just human-centered computing. arXiv:2405.03699.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology* 47:55–130.

Haemmerl, K.; Deiseroth, B.; Schramowski, P.; Libovický, J.; Rothkopf, C.; Fraser, A.; and Kersting, K. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics*, 2137–2156.

Homer, B. A., and Fleming, A. C. 2023. Soldier perspectives on the knowledge, skills, and behaviors of a good teammate. Technical Report Research Note 2023-07, Army Research Institute.

Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2023. AI alignment: A comprehensive survey. arXiv:2310.19852.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023. The empty signifier problem: Towards clearer paradigms for operationalising "alignment"' in large language models. arXiv:2310.02457.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6:83–392.

Li, J.; Consul, S.; Zhou, E.; Wong, J.; Farooqui, N.; Ye, Y.; Manohar, N.; Wei, Z.; Wu, T.; Echols, B.; Zhou, S.; and Diamos, G. 2024. Banishing LLM hallucinations requires rethinking generalization. arXiv:2406.17642.

Lipton, Z. 2024. Alignment is now defined so broadly. https://x.com/zacharylipton/status/1771177444088685045.

Möller, N. 2016. Value uncertainty. In Hansson, S. O., and Hadorn, G. H., eds., *The Argumentative Turn in Policy Analysis: Reasoning About Uncertainty*. Switzerland: Springer. 105–133.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 27730–27744.

Padhi, I.; Natesan Ramamurthy, K.; Sattigeri, P.; Nagireddy, M.; Dognin, P.; and Varshney, K. R. 2024. Value alignment from unstructured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1083–1095.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv:2310.03693.

Schein, C., and Gray, K. 2018. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review* 22(1):32–70.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2024. Evaluating the moral beliefs encoded in LLMs. *Advances in Neural Information Processing Systems* 36.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical harms: Scoping a taxonomy for harm reduction. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 723–741.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large language model alignment: A survey. arXiv:2309.15025.

Shneiderman, B., and Muller, M. 2023. On AI anthropomorphism. https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. A roadmap to pluralistic alignment. In *Proceedings of the International Conference on Machine Learning*.

Sudalairaj, S.; Bhandwaldar, A.; Pareja, A.; Xu, K.; Cox, D. D.; and Srivastava, A. 2024. LAB: Large-scale alignment for chatbots. arXiv:2403.01081.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; Liu, Z.; Liu, Y.; Wang, Y.; Zhang, Z.; Vidgen, B.; Kailkhura, B.; Xiong, C.; Xiao, C.; Li, C.; Xing, E.; Huang, F.; Liu, H.; Ji, H.; Wang, H.; Zhang, H.; Yao, H.; Kellis, M.; Zitnik, M.; Jiang, M.; Bansal, M.; Zou, J.; Pei, J.; Liu, J.; Gao, J.; Han, J.; Zhao, J.; Tang, J.; Wang, J.; Vanschoren, J.; Mitchell, J.; Shu, K.; Xu, K.; Chang, K.-W.; He, L.; Huang, L.; Backes, M.; Gong, N. Z.; Yu, P. S.; Chen, P.-Y.; Gu, Q.; Xu, R.; Ying, R.; Ji, S.; Jana, S.; Chen, T.; Liu, T.; Zhou, T.; Wang, W.; Li, X.; Zhang, X.; Wang, X.; Xie, X.; Chen, X.; Wang, X.; Liu, Y.; Ye, Y.; Cao, Y.; Chen, Y.; and Zhao, Y. 2024. TrustLLM: Trustworthiness in large language models. arXiv:2401.05561.

Varshney, K. R. 2024. Decolonial AI alignment: Openness, viśeṣa-dharma, and including excluded knowledges. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 1467–1481.

Wang, Q., and Goel, A. K. 2022. Mutual theory of mind for human-AI communication. arXiv:2210.03842.

Wang, Q.; Walsh, S.; Si, M.; Kephart, J.; Weisz, J. D.; and Goel, A. K. 2024. Theory of mind in human-AI interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 493.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of risks posed by language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond preferences in AI alignment. *Philosophical Studies*.