LLMs' Pluralistic Compatibility

SEVERIN ENGELMANN*, Cornell Tech, United States
RAJIV MOVVA*, UC Berkeley, United States

Amid growing recognition of the influence of large language models (LLMs) on societies around the world, designers, scholars, and practitioners turn to the development and deployment of *value pluralistic* models. This extended abstract critically assesses emerging approaches to pluralistic alignment in LLMs. We distinguish between two primary strategies: *procedural pluralism*, which embeds pluralistic principles into model development processes, and *behavioral pluralism*, which concerns the values LLMs express in interaction. For each, we examine the underlying normative assumptions and commitments, highlighting tensions between design choices and the demands of pluralism. To meaningfully incorporate pluralism into LLM design, scholars must grapple with its conceptual complexity and contested dimensions. Crucially, this includes clarifying the goals of pluralistic alignment and articulating why pluralism matters for a given application context.

Keywords: Values in Design, Large Language Models, Pluralism

Reference Format:
Severin Engelmann and Rajiv Movva. 2025. LLMs' Pluralistic Compatibility. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction

The accelerating progress of generative AI systems, including data-intensive large language models (LLMs), has stirred both excitement and optimism among AI practitioners and system designers. Millions interact with LLMs for guidance on specific tasks as well as social and parasocial relationships as friends, romantic partners, or mentors [11]. Going back at least to the 1980s, a substantial corpus of works in different disciplinary practices have demonstrated how technological artifacts incorporate values (ethical, political, and so on) and how designers can exercise power over the nature of these values [10, 19]. The recognition that AI development and deployment are informed by values extending beyond mere functionality or efficiency has become established within FAccT scholarship and its associated communities [5, 7, 24].

With the ascendant adoption of LLM applications and products into everyday life, each LLM output expresses and embodies *values*—be it cultural, ethical, or political values. Because LLM outputs influence users in profound ways, LLMs not only reflect values but activate and shape users' values in powerful ways. A recent study by researchers at Anthropic analyzed 308,210 user interactions with their flagship model and found that users engage LLMs for a broad range of purposes, including seeking relationship advice, resolving historical controversies, navigating

*Both authors contributed equally to this research.

Authors' Contact Information: Severin Engelmann, Cornell Tech, New York City, United States; Rajiv Movva, UC Berkeley, Berkeley, United States.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

ethical questions in technology, and making morally sensitive requests [12]. In total, the model articulated 3307 unique value expressions across the sampled conversations, underscoring the extent to which generative AI systems can be understood as *value-encoding* and *value-expressing machines*.

A fundamental challenge lies in determining which values these systems ought to encode and express when interacting with people across various cultural and social contexts worldwide. For example, risk taxonomies highlight how LLMs can reproduce stereotypes and biases that misrepresent or overlook certain viewpoints, identities, and cultural perspectives [25, 26]. In response to the numerous ways LLMs express and shape societal values, recent work has introduced techniques aimed at developing so-called 'pluralistic LLMs' or 'pluralistically aligned LLMs' [9, 14, 21, 22].[1] While these attempts are laudable and warrant continued attention by the AI ethics community as a whole, in particular FAccT and EWAF, it is critical to assess the underlying assumptions of value pluralism operationalized in these initial designs. The goal of this extended abstract is to take a first step into this direction.

2 Approaches to procedural pluralism in LLMs

Pluralistic principles can shape model *development* at all stages of the model development pipeline. Currently, most procedural approaches in value pluralism focus on the fine-tuning stage of model development often referred to as "alignment" [6].[2] Consequently, we here document current approaches to pluralistic model development during alignment. However, we note that the focus on alignment ignores that base models are a site of value considerations and choices, too. For example, developers decide what websites to sample from for pre-training and which ones to blacklist due to undesirable properties.[3]

Broadening standing in RLHF. In Reinforcement Learning from Human Feedback (RLHF), human annotators rank between model-generated pairs of responses R_A and R_B. These *preferences* are used to train a reward model which estimates a score for how much an output will be preferred, and the LLM is fine-tuned to increase this score via RL. If human feedback orientates models' expressed values, naturally, a key question is: whose values should orientate LLMs? "Standing" is typically granted to a group of hired crowdworkers whom model developers deem to produce high-quality feedback on model-generated outputs [2, 20]. A common procedural consideration for pluralistic model development has been to pluralize the demographic and geographic makeup of annotator groups. The rationale is that pluralistic standing ultimately produces models with a wider range of values in different tasks, topics, or functions (e.g., [17]). However, because RLHF eventually aggregates individual human value choices—expressed as preferences on model output alternatives— into numerical, scalar reward signals, it is questionable whether broadening standing translates into pluralistic value expression in LLMs. How normatively

[1] See also the recent Pluralistic Alignment Workshop at NeurIPS 2024: https://neurips.cc/virtual/2024/workshop/84737.

[2] Many leading models do not describe alignment processes in detail; we rely on information available to the research community, which is the same information researchers use to inform their methods for pluralistic alignment.

[3] Such and other choices affect what values models represent and express. Lucy et al. (2024) show that common pre-training filters disproportionately exclude text from non-Western countries and certain social groups [18]. West and Potts (2025) illustrate how unaligned base models are more creative and original in poetry writing and game strategizing compared to aligned models' tendency to produce safe, pleasant, and predictable outputs [27]. Base models embody values that may be traded off against other desiderata during fine-tuning.

valuable is pluralistic standing as a meaningful integration of pluralism when incommensurable (i.e., pluralistic) value choices are eventually aggregated and reduced to a single reward model that fine-tunes the LLM?[4]

A more fundamental philosophical inquiry is required to assess whether a given pluralistic process is legitimate in light of the conceptual commitments underlying a specific notion of value pluralism. While a full treatment of this question exceeds the scope of this abstract, it is worth noting that a strict ontological conception of value pluralism likely disqualifies pluralistic RLHF as legitimate. An ontological conception treats values as structurally incommensurable: that is, values may be compared and debated, but not according to a *shared* measurement scale. Value monism, the conceptual opposite of value pluralism, assumes values are commensurable because they can be subsumed under a single overarching value, such as utility [15, 28]. A more pragmatic conception of value pluralism might deprioritize ontological commitments, shifting the focus instead to whether model outputs reflect values that are appropriate for the norms and expectations of a given context [1]. Different conceptual characteristics of value pluralism specify normative demands and hence inform and justify particular design choices.

Pluralistic value infusion through collectively-developed constitutions. Instead of learning values from preference data, recent approaches attempt to inject LLMs directly with a list of predetermined values [3, 4, 13]. To incorporate a list of values, the LLM is fine-tuned using critiques from a separate LLM (which provides *AI feedback*, as opposed to human feedback). In this case, human input can be provided during value selection, but the encoding of such values is left to the feedback LLM. For example, in constitutional AI, an LLM is prompted to evaluate which of two outputs better aligns with a list of high-level, abstract values, referred to as a "constitution" [3]. Values in the constitution are natural language principles that represent abstract normative commitments such as "ethical and moral commonsense standards," or directives such as "the AI should provide balanced and objective information reflecting all sides of an issue." The authors declare that these values were developed in a mostly arbitrary manner.[5]

More recent approaches use a collective approach in determining constitutional values for LLM fine-tuning. For example, in the *collective* constitutional AI (CCAI) approach, a representative sample of US crowdworkers was recruited to deliberate and hence vote on constitutional values to guide RLAIF [13]. A similar deliberative process that solicits user values is the Sociotechnical Language Model Alignment (STELA; [4]). Researchers recruited participants from four demographic minority groups in order to identify values that might be overlooked by population-wide sampling. Drawing on theories of deliberative democracy, the researchers first facilitated internal deliberations within each group, then identified a shared set of values across groups. Finally, one expert from each group was selected to review and validate the final value set, ensuring that the researchers' formulations remained faithful to the groups' original input. Similar to CCAI, the resulting output is a list of values intended to guide model behavior.

CCAI and STELA add participatory elements to the process that determines what values guide LLM fine-tuning. However, participatory approaches currently neither account for how values are interpreted by the feedback model nor whether the feedback model's values correspond to the values of the collective. In short, these approaches suffer from a limitation of circularity: they rely on an *LLM's* interpretation of values deemed prescriptive for model training. For example, for a feedback model to decide which of two responses is more "balanced and objective,"

[4]For a general critique of the reductive nature of RLHF see [29].

[5]As per Anthropic: "These principles were chosen in a fairly ad hoc and iterative way for research purposes. In the future, we believe such principles should be redeveloped and refined by a larger set of stakeholders…" See [3] for further details.

it must rely on its existing encoding of "balanced and objective." There is no reason to assume that the feedback model's judgments will align with that of the users who developed the constitution.

This disconnect might be termed the *encoding gap*: there is a gap between how the principles are justified in the first place, via public input, and how they are operationalized, by relying on the judgments of LLMs with difficult-to-trace data provenance. The encoding gap raises doubt as to how meaningfully pluralistic participation translates into pluralistic LLMs. Addressing the encoding gap will require more meaningful user participation across the full lifecycle of model development [23]. This means involving users not only during the design phase, but also during deployment. For example, users can work with model developers to evaluate how effectively a model operationalizes agreed-upon principles. Such participation must be iterative: users need to observe how principles are instantiated in practice to determine whether the resulting model behavior aligns with their expectations and values. Recent work in human-computer interaction has begun to address this intention–outcome gap by allowing users to refine and iterate on lists of LLM principles to better achieve their desired results [8]. However, a structural tension remains: deeper forms of participation often come at the cost of scalability, limiting the number of users who can be meaningfully involved.

3 Approaches to behavioral pluralism in LLMs

Another site for value-pluralistic LLM design is LLM *behavior*. Several distinct commitments can be identified. One perspective views pluralistic LLM behavior as an opportunity to empower individual users by aligning with their personal values. In this sense, it may not differ significantly from the goals of personalization that underpin recommender systems in search engines and social media platforms. This *user-centered* pluralism operates on the assumption that value pluralism in LLMs should help individuals and their communities sustain and strengthen their value identities. Under this notion, LLMs situate individuals within distinct value communities and reflect these values back affirmatively in its outputs [16]. Value pluralism is thus accomplished at the societal level where the consolidation of individuals' values leads to a more pluralistic society overall. By placing ultimate normative authority for value pluralism in individuals and their communities (and, by extension, their cultures, traditions, and conventions), user-centered pluralism risk instantiating relativistic rather than pluralistic LLMs [28].

Topic-centered pluralism anchors pluralistic considerations on the conversational topics, tasks, and contexts users choose LLMs to interact with. It mandates an LLM to recognize and represent a variety of perspectives, beliefs, and worldviews. For example, when an LLMs is prompted to converse about controversial topics such as euthanasia, gun control, or social media regulation, topic-centered value pluralism requires an LLM to consider a multiplicity of perspectives with respect to a specific topic. Topic-centered value pluralism is educational, prescribed to expose users to values different to their own.

In one version of topic-centered pluralism, an LLM is instructed to present a variety of different viewpoints and perspectives on a topic by defining an "Overton window," a selection of reasonable but divergent values in its response to the user [22]. This presupposes that an LLM can identify a set of values that are not only relevant and sufficiently distinct, but also normatively reasonable within a given topic, task, and context. The notion of an Overton window, borrowed from political theory, raises important normative questions: who decides which values are reasonable? If the Overton window simply mirrors majority-held views, it remains unclear how such an approach avoids reinforcing dominant norms at the expense of marginalized perspectives. As a partial solution, Sorensen et al. propose using an entailment model to distinguish between "reasonable" and "unreasonable" value

positions [22]. Yet this again begs the question of which normative assumptions underpin the entailment model itself. Moreover, other approaches to pluralism raise parallel concerns: steerable pluralism, which tailors outputs to user preferences, risks reinforcing filter bubbles and echo chambers; distributional pluralism, which grounds responses in value distributions across populations, lacks clear criteria for how those populations are selected. This makes the ethical justification of current approaches to pluralistic modeling fragile.

Both user-centered and topic-centered pluralism ultimately require that LLMs model value variety. That is, they must be capable of representing multiple, often conflicting, human values in a given situation. This demands the development of modeling strategies that not only recognize value variety, but can also explain and balance these values in contextually-appropriate ways. An example of this effort is Kaleido [21], a "value pluralistic model" designed to generate, explain, and assess the relevance and valence (i.e., support, oppose, context-dependent) of multiple human values in morally complex scenarios. Kaleido operates on the premise that if a model can reliably represent value variation in a single context, it can be aligned with broader pluralistic objectives. To realize this, the authors created *ValuePrism*, a dataset comprising 218,000 annotated value judgments across 31,000 ethically charged situations. The situations were generated by crowdworkers, and the value labels were generated by GPT-4. Kaleido was then fine-tuned on these synthetic data.

While Kaleido outperforms GPT-4 on several benchmarks (e.g., Commonsense Norm Bank, ETHICS), the approach raises concerns about inherited biases both from crowdworkers and GPT-4. Notably, this approach provides no justification for how GPT-4 identifies and generates the set of values for each human-authored scenario. This lack of transparency complicates claims about the validity of the value representations encoded in the model. To partially address concerns about representativeness, the authors recruit crowdworkers from marginalized backgrounds and report that these workers validate the pluralistic quality of the ValuePrism dataset to the same extent as non-marginalized groups. Nevertheless, the effectiveness of Kaleido's modular and interpretable architecture depends critically on the comprehensiveness and justification of the underlying datasets. Thus, Kaleido illustrates both the promise and the limitations of pluralistic modeling in LLMs: it demonstrates that modeling value variety is feasible, but also that meaningful behavioral pluralism requires ongoing efforts to expand, validate, and scrutinize the value annotations, datasets, and modeling choices that underpin these systems.

4 Conclusion

The concept of pluralism is often invoked in discussions of AI ethics, but its meaning and implications remain far from settled. Pluralism is not a monolithic ideal but a diverse and contested one, encompassing multiple dimensions. Kekes, for example, distinguishes among ontological, epistemological, and axiological pluralism (concerning what makes life good), as well as anthropological and political forms—each reflects different aspects of human diversity and normative commitment [15]. Before translating pluralist commitments into technical systems, we must ask: what are our main goals in doing so? Why should pluralism matter in the context of LLM design? Normatively speaking, we need to ask: pluralism for what, and for whom? What, exactly, are we trying to pluralize? Without clarity on these foundational questions, pluralistic alignment risks becoming an empty placeholder rather than a meaningful design commitment.

References

[1] Elizabeth Anderson. 1995. *Value in ethics and economics*. Harvard University Press.

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. https://doi.org/10.48550/arXiv.2204.05862 arXiv:2204.05862.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. https://doi.org/10.48550/arXiv.2212.08073 arXiv:2212.08073 [cs].

[4] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports* 14, 1 (March 2024), 6616. https://doi.org/10.1038/s41598-024-56648-4 Publisher: Nature Publishing Group.

[5] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 173–184.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[7] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What people think AI should infer from faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 128–141.

[8] K. J. Kevin Feng, Inyoung Cheong, Quan Ze Chen, and Amy X. Zhang. 2024. Policy Prototyping for LLMs: Pluralistic Alignment via Interactive and Collaborative Policymaking. https://doi.org/10.48550/arXiv.2409.08622 arXiv:2409.08622 version: 1.

[9] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. https://doi.org/10.48550/arXiv.2406.15951 arXiv:2406.15951.

[10] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. 2008. *Embodying values in technology: Theory and practice*. na.

[11] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. https://doi.org/10.48550/arXiv.2404.16244 arXiv:2404.16244 [cs].

[12] Saffron Huang, Esin Durmus, Miles McCain, Kunal Handa, Alex Tamkin, Jerry Hong, Michael Stern, Arushi Somani, Xiuruo Zhang, and Deep Ganguli. 2025. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint arXiv:2504.15236* (2025).

[13] Saffron Huang, Divya Siddarth, and Liane Lovitt. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. (2024).

[14] Atoosa Kasirzadeh. 2024. Plurality of value pluralism and AI value alignment. https://openreview.net/forum?id=AOokh1UYLH

[15] John Kekes. 1996. *The morality of pluralism*. Princeton University Press.

[16] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (April 2024), 383–392. https://doi.org/10.1038/s42256-024-00820-y

[17] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. http://arxiv.org/abs/2404.16019 arXiv:2404.16019 [cs].

[18] Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7393–7420. https://doi.org/10.18653/v1/2024.acl-long.400

[19] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155 arXiv:2203.02155 [cs].

[21] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19937–19947.

[22] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. http://arxiv.org/abs/2402.05070 arXiv:2402.05070 [cs].

[23] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1609–1621.

[24] Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Yuko Ikkatai, Naira Paola Arnez-Jordan, Rose Caleno, Brian Mboya, Shuichiro Higuma, Tilman Hartwig, Hiromi Yokoyama, et al. 2024. Attitudes toward facial analysis AI: A cross-national study comparing Argentina, Kenya, Japan, and the USA. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2273–2301.

[25] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. Using the Veil of Ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences* 120, 18 (May 2023), e2213709120. https://doi.org/10.1073/pnas.2213709120 Publisher: Proceedings of the National Academy of Sciences.

[26] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[27] Peter West and Christopher Potts. 2025. Base Models Beat Aligned Models at Randomness and Creativity. *arXiv preprint arXiv:2505.00047* (2025).

[28] Carla Yumatle. 2015. Pluralism. *The Encyclopedia of Political Thought, First Edition. Edited by Michael T. Gibbons, Published* (2015).

[29] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond Preferences in AI Alignment. https://doi.org/10.48550/arXiv.2408.16984 arXiv:2408.16984 [cs].