Your Model Is Not Neutral—It's Just Well-Socialized

Ananyapam De
TU Clausthal
ananyapam.de@tu-clausthal.de

Benjamin Säfken
TU Clausthal
benjamin.safken@tu-clausthal.de

Abstract

We argue that contemporary instruction-following language models are best understood not as "neutral" calculators over text but as *well-socialized* systems: they internalize, reweight, and strategically present culturally dominant norms and rules. We synthesize evidence spanning pretraining corpora composition, alignment pipelines (RLHF/RLAIF), and evaluation artifacts, and we propose a measurement protocol that distinguishes *norm adherence* from *value neutrality*. Our position reframes bias mitigation: the question is not whether models encode values, but *which* values they surface, to whom, and under what prompting regimes. Our contribution is conceptual and procedural: a vocabulary and checklist, not new metrics or experiments.

1 Introduction

LLMs are trained on vast web corpora that imprint human regularities, including stereotypes and social hierarchies [7, 9]. Subsequent scaling and deployment has foregrounded risks and normative choices [2, 4]. Alignment methods such as RLHF [20] and RLAIF/Constitutional AI [1] further *curate* model behavior toward socially sanctioned responses. The resulting systems often *exhibit* politeness, safety, and fairness while still inheriting corpus-level asymmetries. We label this phenomenon *well-socialization*: optimization that shapes surface behavior to conform to desired norms without rendering the representation space norm-neutral.

Claim. LLMs are not neutral; they are *socialized* by (i) data, (ii) objective and feedback design, and (iii) evaluation/benchmarking conventions. Recognizing this clarifies why debiasing sometimes improves benchmark scores yet leaves downstream disparities intact.

2 From Neutrality to Well-Socialization

Pretraining socialization. Web-scale corpora (e.g., C4) include skewed topical and demographic coverage and blocklist filtering that *systematically* drops minority varieties [9]. Word- and sentence-level associations capture culturally learned stereotypes [7]. These are not incidental noise but stable statistical structure.

Alignment socialization. Instruction tuning with human feedback translates community norms into loss gradients [20]. Constitutional AI replaces or supplements humans with principles, still codifying a normative charter [1]. In both regimes, optimization encourages *performative* adherence to rules (e.g., refusal templates), which can mask representational biases.

Documentation socialization. Model and dataset documentation (model cards and datasheets) make norms explicit and can redirect optimization targets [17, 10]. This is constructive socialization: it steers practice toward transparency and stakeholder considerations.

Preprint.

3 Evidence: Values In, Values Out

Corpus effects. C4 analysis shows content source skew and blocklist-induced deletions affecting marginalized dialects [9]. At the representation layer, canonical results (WEAT) exhibit human-like biases learned from unlabeled text [7].

Behavior under bias probes. CrowS-Pairs and StereoSet demonstrate persistent stereotypical associations in strong LMs [19, 18]. In toxicity detection pipelines, dialectal bias produces higher false positives on AAE [23]; analogous mechanisms can arise in generative toxicity proxies.

Alignment trade-offs. RLHF improves helpfulness and reduces overt toxicity [20] while leaving latent associations that reappear under adversarial or distribution-shifted prompts. Constitutional AI encodes a principled rulebook but *selects* which rules govern surface behavior [1]. Thus, alignment changes *which* values are expressed, not whether values are present.

Scholarly meta-evidence. Surveys emphasize that "bias" work is intrinsically normative and must ground harms and stakeholders [4]. FAccT meta-analyses document how ML research itself encodes community values [3]. Broader STS analyses trace datasets, labor, and resource externalities [8]. We operationalize well-socialization along three measurable axes, adding concrete metrics and multiple corroborating references for each.

4 Operationalizing Well-Socialization

Representational Associations. (i) Compute classic WEAT/SEAT scores on base and aligned LMs using [7] and [16] templates; for contextual encoders or LLMs, use masked/probe variants from [14]. (ii) Audit residual bias after debiasing or alignment ("lipstick on a pig") by replicating [12] and inspecting *indirect* bias in geometry. (iii) Test word-level association shifts versus sentence-level behavior by comparing pre-/post-alignment embeddings [6] and generation on stereotype prompts.

Surface Norm Adherence. (i) Quantify policy conformance and refusal propensities under instruction-following regimes using preference-optimized models [25, 20] and constitutions [1]. Report refusal rate, helpfulness/harmlessness Elo, and policy-consistency. (ii) Stress-test with toxicity and safety suites (toxicity rate, conditional toxicity, detox controllability) using RealToxicityPrompts and multi-metric dashboards like HELM [11, 15]. (iii) Probe brittleness via automated red teaming and universal jailbreak suffixes; measure attack success rate (ASR) and robustness under prompt perturbations [22, 26].

Context Sensitivity. (i) Dialect and register audits: evaluate toxicity and moderation outcomes on African-American English (AAE) and other varieties [5, 23]. (ii) Cross-modality spillover: check text behavior conditioned on ASR transcripts from demographically diverse speakers (error-driven bias) [13]. (iii) Identity-controlled templates: use HOLISTICBIAS descriptors across demographic axes and QA bias benchmarks like BBQ to measure identity-conditioned preference shifts, refusal asymmetries, and harms [24, 21].

A model is "well-socialized" if surface norm adherence is high while representational associations remains non-neutral or context sensitivity reveals conditional disparities. Reporting should include all three, with stratified slices and uncertainty.

5 Implications

Benchmarking. Include paired probes: (A) value expression under neutral prompts; (B) value expression under adversarial/realistic prompts; (C) robustness across sociolinguistic varieties [19, 23].

Design. Treat alignment as *policy learning*. Make the normative charter explicit (constitution/model card) [17, 10]. Audit feedback provider pools; diversify principles and rater demographics.

Governance. Document data lineage and filtering, publish datasheets/model cards with stakeholder input, and report failure cases and trade-offs rather than implying neutrality [10, 17].

6 Conclusion

Neutrality is not the default; socialization is. LLMs learn cultural regularities from pretraining and then are coached to present rule-following behaviors via alignment. Recognizing this reframes fairness from "removing bias" to *governing which values are encoded and expressed*, with transparent charters, pluralistic feedback, and multi-layered evaluation.

References

[1] 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. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623. ACM, 2021.

[3] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Rida Qadri, and Inioluwa Deborah Raji. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, 2022.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[5] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of EMNLP*, pages 1119–1130, 2016.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 2016.

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[8] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, CT, 2021.

[9] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305. Association for Computational Linguistics, 2021.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[11] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, pages 3356–3369, 2020.

[12] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614, 2019.

[13] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Mina Quartey, Zayd Mengesha, Chelsea Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[14] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in NLP*, pages 166–172, 2019.

[15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[16] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social bias in sentence encoders. In *Proceedings of NAACL-HLT*, pages 622–628, Minneapolis, Minnesota, 2019.

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229. ACM, 2019.

[18] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2021.

[19] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020.

[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. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[21] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut Zhang, James Michaelov, and Adina Williams. Bbq: A hand-built bias benchmark for question answering. In *Proceedings of NAACL*, 2022.

[22] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, Gabriel Goh, et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[23] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[24] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of EMNLP*, pages 9180–9211, 2022.

[25] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020.

[26] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.