Model Plurality: A Taxonomy for Pluralistic AI

Christina Lu
University of Oxford
christina.lu@cs.ox.ac.uk

Max Van Kleek
University of Oxford
emax@cs.ox.ac.uk

Abstract

This position paper argues that the project of pluralistic AI should be expanded from diversifying the values of individual models towards a systemic pluralism that allows for new values to emerge. First, we examine the dangers of homogeneity within the existing landscape of public-facing machine learning models. Beyond uplifting certain values over others, models have the potential to reinforce arbitrary biases and homogenize the very ontologies with which we think. We argue for model plurality—structurally embedding multiplicity into every level of model development and deployment via technical strategies and socioeconomic incentives—as a design method for addressing these dangers and creating models with meaningful difference. Finally, we provide a taxonomy of model plurality that organizes the production pipeline into areas of intervention: data, architecture, fine-tuning, and ecosystem. At each level, we analyze incentives that maintain the status quo of homogeneity, what benefits plurality could produce, and sociotechnical approaches for instantiating a more comprehensive plurality in that domain. Model plurality may not only create less biased and more robust models, but also the conditions for the ongoing evolution of human values.

1 Introduction

As public use of generative models grows, and with it their ability to color societal values, the political valence of model output becomes hotly contested by either side of the aisle. Current discussion of pluralistic AI focuses on values plurality: how can a large language model (LLM) be aligned to a spread of Overton-reasonable values rather than biased towards a particular worldview [24]? However, while models can be steered towards different behaviors at a surface level, the structure of knowledge that underpins their values remains constant. A single model can only contain a single ontology—the categories, concepts, and their interrelations from which its worldview is derived. The dominance of any one model has the potential to homogenize ontologies and hence values across society writ large.

What is required, then, is a form of pluralistic AI that is not limited to individual models alone. This position paper argues for an expanded understanding of model plurality that embeds multiplicity into every step of development and deployment. From the structure of the model to the structure of the market, plurality emerges from the coexistence of meaningful difference. This approach does not discard the project of aligning singular models to diverse perspectives but treats it as a sub-problem of a larger endeavor. Re-orienting from values pluralism to a structural pluralism makes comprehensively countering the dangers of homogeneity possible. Minimally, it could create more robust and less biased models; maximally, it could allow for new ontologies to become thinkable. We present a taxonomy that gives structure to the areas where multiplicity could be instantiated: data, architecture, fine-tuning, ecosystem. This taxonomy for model plurality may help create an landscape of machine learning models that is not simply plural, but also breeds the potential for a robust and self-perpetuating diversity.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Background

Algorithmic monoculture. Literature on homogeneity in algorithms has generally focused on algorithmic decision-making systems. Previously, those seeking jobs or loans would face individual humans capable of making biased or arbitrary decisions, but regardless, each unique decision-maker would employ their own unique criteria. When many human decision-makers are replaced by a sector-wide algorithm, such as a candidate screening software used by more than a third of Fortune 100 companies [27], a singular set of criteria codified by the algorithm becomes broadly applied. Creel and Hellman [6] identify the ethical harms of these algorithms not in the arbitrariness of their decisions but in the systematicity of their reach.

Beyond hurting specific individuals, algorithmic monoculture worsens overall performance across a sector. Kleinberg and Raghavan [17] present a mathematical proof for an unintuitive Braess' paradox where the collective quality of decision-making declines when multiple firms use the same algorithm, even if the algorithm makes more accurate decisions than individual firms. Pivoting to model architectures in deep learning, Fishman and Hancox-Li [13] argue that the simplicity, parsimony, and generality of unified model architectures does not outweigh its epistemic and ethical risks including path dependency, increased black-boxing, and centralization of power. Technical risks of homogeneity are also well-trodden on the cybersecurity front: monoculture breeds single points of failure that make the entire ecosystem susceptible to individual bugs or exploits [8].

More recently, the phenomenon of model collapse—generative AI models deteriorating in quality when recursively trained on generated data—is also traceable to the homogeneity of these models, with their content repetitively flattened from the original distribution by lossy compression and smoothed averages [23]. Ultimately, the systematizing force that a dominant, general-purpose machine learning model is poised to wield presents both a social and technical problem.

Machine ontologies. The primary concern of pluralistic alignment is to align AI with a multiplicity of human values, contending with the complex and often incommensurate dimensions of human moral primitives. While values are the most tangible referent for grappling with our differences, we are concerned with that which underlies them: ontology. Referring to the categories and relations between them which make up a given worldview, ontologies are integral to how any values are understood. There is no single universal ontology for defining the world but rather a multitude of both individual and collective ones, that are constantly updated, expanded, and pruned. As theorized by Denizhan [7], diversity of ontologies is important because it the constant interaction and synthesis of ontological difference that allows for the ongoing evolution of semantic concepts such as values.

We argue that a single machine learning model is only capable of maintaining a single ontology. Output valence can be colored by few-shot learning on system prompts, but the underlying parameter weights that make up the model's knowledge are static. While parameters cannot be directly translated to an ontology, research in mechanistic interpretability looks at neuron activations to better understand the relationship between the latent space and semantic meaning [2]. Identifiable patterns in neuron activations appear for high-level semantic concepts such as categories, hierarchical nesting [20], or contradicting true-false statement pairs [4]. This suggests that while rich concepts can be embedded within the model, there remains a single geometry organizing these concepts. It follows that training with similar datasets, architectures, and methods would produce models that contain similar geometries within the latent space and hence, ontologies [31]. Thus, even the current variety of public-facing generative LLMs produce similar machine ontologies, exposing and entraining the diversity of human ontologies to their pervasive, homogenizing influence.

Model plurality. Models with similar training pipelines produce similar ontologies. Model plurality must address this homogenization of machine ontologies and preserve the foundation that makes all present and future values thinkable. This will require intervening at every step of development, from training data to the architecture to the fine-tuning process, to deliberately differentiate models from one another and embed multiplicity where possible. Expanding beyond individual models, model plurality must also intervene within the wider ecosystem of deployment. It must counter the market forces which push users toward a single product and ensure that use of different models is distributed across society. Creating a more comprehensive model plurality will require novel political-economic incentives and technical interventions alike. We specifically advocate for a robust, diverse landscape of machine learning models that constantly reproduce the conditions for an ongoing plurality—which may lie in a future of more bespoke yet compositional models.

3 A Taxonomy of Model Plurality

In this section, we present a multi-tier taxonomy of model plurality that describes where multiplicity could be inserted across the production pipeline. For each tier, we touch on the incentives for existing homogeneity, discuss possible ways to instantiate multiplicity, and speculate on the benefits such multiplicity could provide. By providing structure to this wide problem space, we hope to make sociotechnical intervention more feasible.

3.1 Data

The data tier refers to the corpus on which the model is trained. Due to the current paradigm of unsupervised learning, where models derive general capabilities from training on the largest possible dataset, scraping the internet for public content becomes the only method for satiating this demand [3]. Thus, large models tend toward learning to approximate the same content, in the same modalities available on the internet: text, image, video. There is a vast landscape of information out there that is not privy to machine learning models because the data is not publicized, digitized, recorded, or even measured.

Our recommendations for model plurality here are twofold: diversifying the data models are trained on beyond the detritus of the internet and exploring alternate model architectures that require less data. Expanding the training corpus of models is limited by concerns around protecting data privacy and value [11]. However, privacy-enhancing technologies may help move beyond a binary of public versus private [29, 22], allowing models to be trained on currently unseen data to produce different and specific worldviews. At the same time, governance protocols by Spawning [25] intend to address IP concerns through consent-based scraping and data compensation frameworks. Further development in these areas could allow the breadth of models to diverge from learning a homogenized dataset and instead speciate thanks to siloed, privatized data. The second vector for plurality involves developing model architectures that do not require massive amounts of data or compute in order to achieve generalized ability (and indeed, questioning the goal of generality itself); this may include reusing techniques for under-resourced languages but also recomposing models from smaller ones, as will be discussed in the following section. Nurturing multiplicity in data could address some of the current marginalizing tendencies within massive general models that converge on a mean and erase outliers.

3.2 Architecture

The architecture tier bears the most resemblance to existing work in pluralistic AI as it seeks to embed multiple worldviews within the structure of the model itself. Current model architectures within public-facing generative models are generally identical transformers and diffusion models that are able to take advantage of massive amounts of data [1, 28]. There is little diversity within the landscape as companies prefer low risk investment into scaling compute and data rather than high risk alternative architecture exploration [26].

However, such exploration could simply involve duplicating parts of models, such that each division can maintain their own distinctions. Whether these parallel processes are models, agents, or even simulated debate, these subdomains of difference can then be aggregated for the final result. Where segmentation occurs and how consensus is attained are the levers to toggle in pluralistic architecture experimentation. In this section we focus on multiplicities contained within a single architecture, defined as trained under a shared objective. Du et al. [10] saw mathematical and strategic reasoning improve when multiple instances of LLMs debated one another. Li et al. [18] demonstrated that even asking language models to simulate multiple agents was enough for performance improvement across similar benchmarks. While these techniques sample the possibility space produced by homogenous LLMs with the same architecture, other frameworks fuse, mediate, and integrate heterogeneous LLMs with different architectures altogether [16, 30]. Moving a level deeper than agent-level pluralism, Du et al. [9] discuss compositing large, complex generative models from small, simpler ones. They argue that this is both data and compute-efficient and better at generalizing to previously unseen distributions. Adding multiplicity directly into the architecture of models increases robustness and performance, and further research could explore compositing alternative ontologies together.

3.3 Fine-tuning

The fine-tuning tier refers to various techniques that come after pretraining that still modify model parameters to adjust behavior. This section focuses on reinforcement learning from human feedback (RLHF), which most overtly focuses on alignment and is thus instrumental in shaping the final output of public-facing generative models. Vanilla RLHF fundamentally eliminates plurality by optimizing for inter-annotator agreement, discarding the nuances of disagreement in favor of convergence around the mean [33]. The annotators themselves are primarily recruited from Anglophone countries in the Global South [5]; the globalized data annotation industry exploits income differentials, resulting in pools of homogenous annotators from the same regions. Models with planet-spanning impact thus take on regional inflections [14].

For the technical problem, ongoing research in multi-objective RLHF that ensemble model parameters is promising [15]. This often involves contending with trade-offs along the Pareto frontier and interpolating parameters from multiple independent networks with different preferences [21]. Other techniques aim to be RL-free, instead directly combining language modeling with the reward modeling of diverse preferences [32]. The latter problem, on the other hand, is borne of global macroeconomic dynamics and requires large-scale policy interventions and adjustments of market incentives. Annotators should be chosen more deliberately, ideally by sampling from target user populations for each particular model. Further transparency around the process of selecting annotators and the annotation objectives they impart onto models is necessary, perhaps through frameworks similar to Model Cards [19]. All models will be biased in some way; the purpose is to make decisions around biases explicit.

3.4 Ecosystem

The ecosystem tier moves beyond individual models to finally consider instantiating plurality within the landscape of models available to the public. Models produced by different companies are roughly similar due to the homogenization of data, model architecture, and training method; most differences come from particularities of the alignment process and system prompt. This section consists of a speculative sketch of ecosystem plurality as it requires fundamentally reimagining how machine learning models are trained, deployed, and commoditized. No single model should dominate the landscape but even the way models are conceived should be reorientated entirely: not wholly trained by a single institution but rather composed from smaller, specific parts that are available to compose and recompose into bespoke models. This is necessary for an ecosystem that is not simply plural in an instance but contains a self-perpetuating, dynamic plurality with mechanisms that exponentially explode the space of possible models.

What if larger models were actually a patchwork of individually-tailored ones? Work in modular plurality in particular demonstrates aligning submodels to underrepresented communities within a setup of multi-LLM collaboration [12]. Building on the compositional generative model framework of Du et al. [9], a future pluralistic ecosystem might consist of wiring models that have learned specific probability distributions with one another to create larger model assemblages. Such assemblages could be spun up ad-hoc for task-specific needs and also serve as customizable, personalized agents for individual human users. Wiring model assemblages of individual humans together could also create super-assemblages of fluid communities that resolve collective decisions and wield collective bargaining for populations. A compositional approach to plurality naturally produces a rich, dynamic ecosystem of machine learning models.

4 Conclusion

In this paper, we have argued for conceiving of model plurality as structural and emergent, of which diversifying the values of an individual model is but a subproblem. We present a taxonomy of this reorientation of model plurality, giving structure to where multiplicity could be seeded within the pipeline of model production: via data, architecture, fine-tuning, and the ecosystem. Model plurality offers both technical and ethical advantages to the existing paradigm of machine learning research. Technically, it creates more robust models that are less susceptible to single points of failure. Ethically, it expands the ontologies that are captured within models and available to the general population at large, allowing for diversity and dynamic possibility in not simply the ontologies themselves but also the morals and values that they give meaning to.

Acknowledgments and Disclosure of Funding

The authors did not receive direct funding for this work. C.L. would like to thank Kellogg College at the University of Oxford for travel support to attend NeuRIPS.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2023. Accessed: 2024-09-09.

[2] Anthropic. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, May 2024. Accessed: 2024-09-09.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024.

[5] Benedetta Catanzariti, Srravya Chandhiramowuli, Suha Mohamed, Sarayu Natarajan, Shantanu Prabhat, Noopur Raval, Alex S. Taylor, and Ding Wang. The global labours of ai and data intensive systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21 Companion, page 319–322, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

[7] Yagmur Denizhan. Intelligence as a border activity between the modelled and the unmodelled. *Angelaki*, 28(3):25–37, 2023.

[8] Kevin Dooley. *Designing Large Scale Lans*. Help for network designers. O'Reilly Media, 2002.

[9] Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need, 2024.

[10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[11] European Parliament and Council of the European Union. Directive (eu) 2019/790 of the european parliament and of the council of 17 april 2019 on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec (text with eea relevance), Apr 2019. Accessed: 2024-09-09.

[12] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration, 2024.

[13] Nic Fishman and Leif Hancox-Li. Should attention be all we need? the epistemic and ethical implications of unification in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, volume 11 of *FAccT '22*, page 1516–1527. ACM, June 2022.

[14] Alex Hern. Techscape: How cheap, outsourced labour in africa is shaping ai. *The Guardian*, Apr 2024. Accessed: 2024-09-09.

[15] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023.

[16] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023.

[17] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[18] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need, 2024.

[19] 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. ACM, January 2019.

[20] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models, 2024.

[21] Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2023.

[22] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning, 2018.

[23] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, Jul 2024.

[24] 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. A roadmap to pluralistic alignment, 2024.

[25] Spawning AI. Spawning ai: Tools for artists in the age of ai, n.d. Accessed: 2024-09-09.

[26] Nick Srnicek. *Data, Compute, Labour*, pages 241–261. The MIT Press, May 2022.

[27] Jennifer Strong. Podcast: Want a job? the ai will see you now. MIT Technology Review Podcast, Jul 2021. Narrated by Jennifer Strong, Accessed: 2024-09-09.

[28] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

[29] Andrew Trask, Emma Bluemke, Teddy Collins, Ben Garfinkel Eric Drexler, Claudia Ghezzou Cuervas-Mons, Iason Gabriel, Allan Dafoe, and William Isaac. Beyond privacy trade-offs with structured transparency, 2024.

[30] Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models, 2024.

[31] M. Watkins. What's up with all the non-mormons? weirdly specific universalities across llms, Apr 2024. Accessed: 2024-09-09.

[32] Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization, 2024.

[33] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.