Extended Abstract: Model Multiplicity for Responsible AI

Prakhar Ganesh

McGill University & Mila
prakhar.ganesh@mila.quebec

Motivation

Machine learning has experienced a remarkable rise, with highly sophisticated over-parameterized models leading the way (Yuan 2023). Consequently, these cutting-edge models find application across diverse domains. Their increasing deployment has sparked concerns about their real-world impact. This includes patterns of biased treatment (fairness) (Mehrabi et al. 2021), brittleness to distribution shifts (robustness) (Subbaswamy, Adams, and Saria 2021), information leakage risks (privacy) (Shokri et al. 2017; Carlini et al. 2023), and vulnerability to adversaries (security) (Szegedy et al. 2013), among others, studied under the umbrella of responsible AI.

A crucial aspect of building responsible AI models is the idea of model multiplicity (Black, Raghavan, and Barocas 2022; D'Amour et al. 2022). Born out of overparameterization, multiplicity is the existence of multiple models that perform similarly well on a specific task, despite learning different underlying functions. As these models learn different mappings of the training data, they tend to have diverse trustworthy behaviour despite similar accuracy, further exacerbated for unseen metrics (Ganesh 2024).

If managed well, model multiplicity gives us the freedom to prioritize several metrics, including those associated with responsible AI, and select the best models to minimize harm (Black, Raghavan, and Barocas 2022; Ganesh et al. 2023; Black et al. 2024). However, the existence of multiplicity also marks the unavoidable presence of arbitrariness in model selection that can impact individual-level decisions, necessitating a broader discussion on the role and expectations of AI decision makers in our society (Jain, Creel, and Wilson; Creel and Hellman 2022; Paes et al. 2023; Gomez et al. 2024; Kulynych et al. 2023).

Furthermore, any investigation in multiplicity is a study of the learning dynamics of AI models, i.e. how changes in the optimization impact the model. By exploring this relationship between multiplicity and learning dynamics, we can improve our understanding of the underlying optimization and create powerful steerable models (Ganesh et al. 2023). I aspire to shape a brighter and more responsible future for all by advocating for responsible model development.

Background

Model Multiplicity Multiplicity, as defined by Marx, Calmon, and Ustun (2020), is the occurrence of conflicting predictions across competing models, i.e., models with comparable performance, rooted in overparameterized AI pipelines (D'Amour et al. 2022). While the phenomenon of multiplicity in prediction problems is not new (Nelder and Wedderburn 1972; Breiman 2001), recent research under the umbrella of *model multiplicity* (Black, Raghavan, and Barocas 2022) has broadened the discourse to its real-world impact and potential in advancing responsible AI practices. This includes work on leveraging multiplicity to select better models (Black, Raghavan, and Barocas 2022; Semenova, Rudin, and Parr 2022; Ganesh 2024), studying the trends of multiplicity with fairness and privacy (Ali, Lahoti, and Gummadi 2021; Kulynych et al. 2023; Long et al. 2023; Ganesh et al. 2023; Cooper et al. 2024), better quantification and efficient assessment of associated risks (Heljakka et al. 2022; Paes et al. 2023; Watson-Daniels et al. 2023; Watson-Daniels, Parkes, and Ustun 2023; Hsu et al. 2024; Hsu and Calmon 2022), and raising ethical and legal concerns on the implications of multiplicity and arbitrariness within existing AI ecosystems (Black, Raghavan, and Barocas 2022; Creel and Hellman 2022; Black et al. 2024; Gomez et al. 2024).

Multiplicity and Responsible AI Several recent works have studied the interplay between multiplicity and other responsible AI metrics. Kulynych et al. (2023) showed that improvements in privacy come at a cost of multiplicity and Long et al. (2023) found a similar cost to achieve group fairness. Ali, Lahoti, and Gummadi (2021) showed that addressing unfairness only in examples with high multiplicity effectively eliminates biases and Cooper et al. (2024) showed similarly that abstaining from highly uncertain predictions exhibits a noticeable improvement in group fairness.

Existing literature has also explored the implications of arbitrariness itself, and its integration into the broader landscape of responsible AI. Several studies have shown a disparity in multiplicity across demographics (Cooper et al. 2024; Gomez et al. 2024; Ganesh et al. 2023), thereby disproportionately impacting certain individuals. Further arguments on the impact of arbitrariness on individuals have also been given using moral discussions (Creel and Hellman 2022) and human rights concerns (Gomez et al. 2024).

Methodology

A fundamental question in AI is: *how do models learn?* Despite numerous attempts, our understanding of learning, especially in neural networks, remains rudimentary and fragmented. Interestingly, exploring multiplicity is deeply entangled with the question of learning. Model multiplicity can give us the tools to investigate the learning dynamics of AI models, ultimately improving our understanding of the large-scale models that are penetrating our daily lives. *My research focuses on investigating model multiplicity in AI, leveraging its potential in deploying responsible models and providing a structure to understand their learning dynamics.*

Quantifying Multiplicity As a first step, it's imperative to start by ensuring appropriate quantification of multiplicity and its associated harms. Existing literature focuses on the case of predictive multiplicity and lacks an appropriate discussion of other harms. Even when different models under multiplicity produce the same prediction, the arbitrariness of model selection can still introduce harms, for instance, in the form of changing privacy risks, robustness behaviour, or security concerns, among others. Thus, while predictive multiplicity is fundamental, it is important to recognize that the adverse effects of multiplicity extend beyond predictions. Consequently, further exploration is needed, and we plan to work on a deeper analysis of various individual concerns due to the arbitrariness introduced by multiplicity.

On a separate note, even when quantifying predictive multiplicity, existing methods are impractical and costly. This is because, despite the existence of various metrics for measuring multiplicity in the literature (Cooper et al. 2024; Kulynych et al. 2023; Black, Raghavan, and Barocas 2022; Hsu and Calmon 2022; Long et al. 2023; Marx, Calmon, and Ustun 2020; Heljakka et al. 2022), these metrics all share a commonality—they rely on training an ensemble of models. The burden of training multiple copies of a model can be unreasonable for those utilizing such models, thus posing a significant barrier to auditing model multiplicity. A recent work by Hsu et al. (2024) broke this trend and showed the advantages of using Monte Carlo dropout as an efficient alternative to training multiple models for measuring model multiplicity. This opens up a promising avenue for future research: understanding the relationship between various measures of model instability, such as predictive uncertainty, distance to the decision boundary, loss curvature, gradient norms, etc., and how they connect with model multiplicity.

Benchmarking Models Under Multiplicity Model multiplicity introduces a significant challenge in benchmarking models. For instance, several studies have questioned the efficacy of existing bias mitigation methods (Baldini et al. 2022; Sellam et al. 2021), revealing susceptibility to factors like random seeds. More specifically, we studied the impact of various forms of randomness on model bias in our recent work (Ganesh et al. 2023) and highlighted the dominant impact of training data order on group fairness. Recently, we also showed a similar concern of variance across several different responsible AI metrics (Ganesh 2024), proposing the use of *multiplicity sheets* when benchmarking models to deal with the concerns of model multiplicity.

Despite recent advancements, we still lack a holistic understanding of how to mitigate and manage multiplicity. Notably, we showed in our work that even the most common recommendation in the literature–model selection based on appropriate constraints–can cause overfitting and may not generalize to unseen settings (Ganesh 2024). Improving our understanding of the origins of multiplicity thus becomes imperative to tackling multiplicity in the real world.

Steerable Models When exploiting multiplicity, choosing the best model from similarly performing models can be a good start, but it doesn't scale. Moreover, the likelihood of finding *black swans*, i.e., exceptionally good models, across a random set of models is always quite low (Hsu and Calmon 2022). However, rather than randomly training multiple models, one can use the insights from multiplicity to create locally steerable models. For instance, referring back to our work on the impact of training data order on model fairness (Ganesh et al. 2023), we used the dominant impact of the latest gradient updates to create custom data orders that can manipulate model fairness as desired. Thus, one can avoid searching across models, and instead directly optimize an existing model towards the required behaviour, allowing us to create responsible and safely deployable models.

Studying the Tradeoffs We discussed how multiplicity can potentially allow the selection of models that enhance multiple auxiliary metrics. However, not all metrics can be improved simultaneously. While certain responsible AI metrics can be improved in harmony, others are naturally in conflict with each other. For instance, to reduce bias in AI, the model needs to memorize data from minority groups, which compromises their privacy (Chang and Shokri 2021). Conversely, enforcing privacy prevents the model from relying too heavily on memorization, which can introduce biases into the model (Fioretto et al. 2022). Understanding these trade-offs is crucial in narrowing down our search space for multiplicity and can help us better exploit the multiplicity of various metrics when creating steerable models. In our future works, we aim to study these trade-offs by connecting them fundamentally to memorization in neural networks.

Legal Requirements and Multiplicity The concerns regarding model multiplicity go beyond existing metrics. Several laws across the globe require that automated decision-makers provide a meaningful explanation to affected individuals of how and why the decision was made. However, how can one justify and meaningfully explain an automated decision if there exist many other models with the same performance that would have made a different decision? There is certain arbitrariness in model selection which creates a conflict between what the law requires and the nature of multiplicity in AI. We are currently collaborating with domain experts in an interdisciplinary attempt to identify the technical challenges in the current state-of-the-art of model multiplicity to meet legal requirements and the legal gap between current law and the implications of arbitrariness in model selection. The goal is to find common grounds either at the model level or through updates to the legal language.

References

Ali, J.; Lahoti, P.; and Gummadi, K. P. 2021. Accounting for model uncertainty in algorithmic discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 336–345.

Baldini, I.; Wei, D.; Ramamurthy, K. N.; Singh, M.; and Yurochkin, M. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2245–2262.

Black, E.; Koepke, L.; Kim, P.; Barocas, S.; and Hsu, M. 2024. The Legal Duty to Search for Less Discriminatory Algorithms. *arXiv preprint arXiv:2406.06817*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT*.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–231.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *USENIX*.

Chang, H.; and Shokri, R. 2021. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 292–303. IEEE.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22004–22012.

Creel, K.; and Hellman, D. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1): 26–43.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *JMLR*, 23(1): 10237–10297.

Fioretto, F.; Tran, C.; Van Hentenryck, P.; and Zhu, K. 2022. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv:2202.08187*.

Ganesh, P. 2024. An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification. In *2024 IEEE/CVF WACV*. IEEE.

Ganesh, P.; Chang, H.; Strobel, M.; and Shokri, R. 2023. On The Impact of Machine Learning Randomness on Group Fairness. In *FAccT*.

Gomez, J. F.; Machado, C. V.; Paes, L. M.; and Calmon, F. P. 2024. Algorithmic Arbitrariness in Content Moderation. *arXiv preprint arXiv:2402.16979*.

Heljakka, A.; Trapp, M.; Kannala, J.; and Solin, A. 2022. Disentangling model multiplicity in deep learning. *arXiv preprint arXiv:2206.08890*.

Hsu, H.; and Calmon, F. 2022. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35: 28988–29000.

Hsu, H.; Li, G.; Hu, S.; et al. 2024. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. *arXiv preprint arXiv:2402.00728*.

Jain, S.; Creel, K.; and Wilson, A. C. ???? Position: Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. In *Forty-first International Conference on Machine Learning*.

Kulynych, B.; Hsu, H.; Troncoso, C.; and Calmon, F. P. 2023. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1609–1623.

Long, C.; Hsu, H.; Alghamdi, W.; and Calmon, F. 2023. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive Multiplicity in Classification. *Proceedings of the 37th International Conference on Machine Learning, PMLR*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nelder, J. A.; and Wedderburn, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3): 370–384.

Paes, L. M.; Cruz, R.; Calmon, F. P.; and Diaz, M. 2023. On the Inevitability of the Rashomon Effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*, 549–554. IEEE.

Sellam, T.; Yadlowsky, S.; Tenney, I.; Wei, J.; Saphra, N.; D'Amour, A.; Linzen, T.; Bastings, J.; Turc, I. R.; Eisenstein, J.; et al. 2021. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *International Conference on Learning Representations*.

Semenova, L.; Rudin, C.; and Parr, R. 2022. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Subbaswamy, A.; Adams, R.; and Saria, S. 2021. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*, 2611–2619. PMLR.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv:1312.6199*.

Watson-Daniels, J.; Barocas, S.; Hofman, J. M.; and Chouldechova, A. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 297–311.

Watson-Daniels, J.; Parkes, D. C.; and Ustun, B. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10306–10314.

Yuan, Y. 2023. On the power of foundation models. In *ICML*, 40519–40530. PMLR.