Mila

16-10-2024

The Curious Case of Arbitrariness in Machine Learning

Prakhar Ganesh

Responsible AI Week

Rashomon (1950)

Akira Kurosawa

Mila

Rashomon (1950)

The rape of a bride and the murder of her samurai husband are recalled from the perspectives of a bandit, the bride, the samurai's ghost and a woodcutter.

- IMDb

Rashomon (1950)

The movie underlines the subjectivity of experience and self-interested advocacy of every individual, rather than an objective truth (or our ability to find that objective truth).

Rashomon (1950)

The movie underlines the subjectivity of experience and self-interested advocacy of every individual, rather than an objective truth (or our ability to find that objective truth).

Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect

Rashomon effect is "*a combination of a **difference of perspective** and **equally plausible accounts**, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Mila

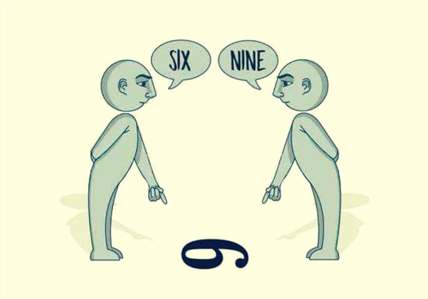
Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the **absence of evidence to elevate one above others**, with the **inability to disqualify any particular version** of the truth...*"

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"



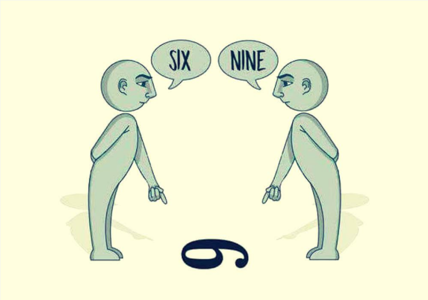
https://classicallyeducated.wordpress.com/2020/05/19/ambrose-bierce-by-way-of-the-rashomon-effect/

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

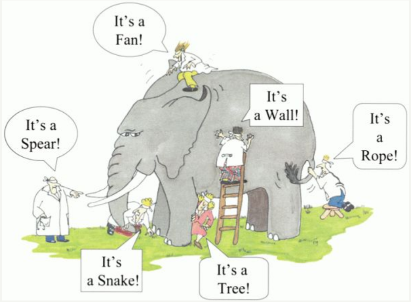
Mila

Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"



https://classicallyeducated.wordpress.com/2020/05/19/ambrose-bierce-by-way-of-the-rashomon-effect/



https://medium.com/stotle-inc/rashomon-effect-lessons-for-building-effective-bi-dashboards-1b484b3137e9

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Mila

Rashomon Effect in Statistical Modeling

Rashomon Effect in Statistical Modeling

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Breiman, L. (2001). *Statistical modeling: The two cultures*. Statistical science, 16(3), 199-231.

Mila

Rashomon Effect in Statistical Modeling

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"What I call the Rashomon Effect is that there is often a multitude of different descriptions (equations f(x)) in a class of functions giving about the same minimum error rate."

Breiman, L. (2001). *Statistical modeling: The two cultures*. Statistical science, 16(3), 199-231.

Rashomon Effect in Statistical Modeling

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

"What I call the Rashomon Effect is that there is often a multitude of different descriptions (equations f(x)) in a class of functions giving about the same minimum error rate."

"Usually we pick the one with the lowest test error. But there may be (and generally are) many equations that have [loss] within 1.0% of the lowest [error]. Which one is better? The problem is that each one tells a different story..."

Breiman, L. (2001). *Statistical modeling: The two cultures*. Statistical science, 16(3), 199-231.

From Rashomon Effect to Multiplicity

From Rashomon Effect to Multiplicity

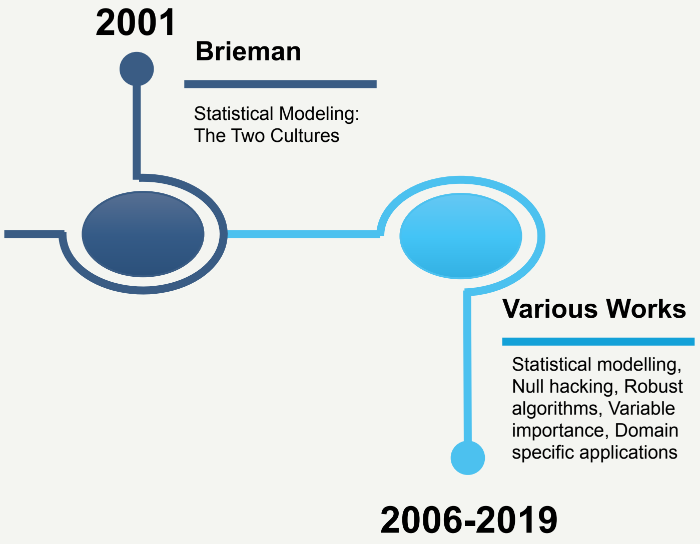
2001

Brieman

Statistical Modeling:
The Two Cultures

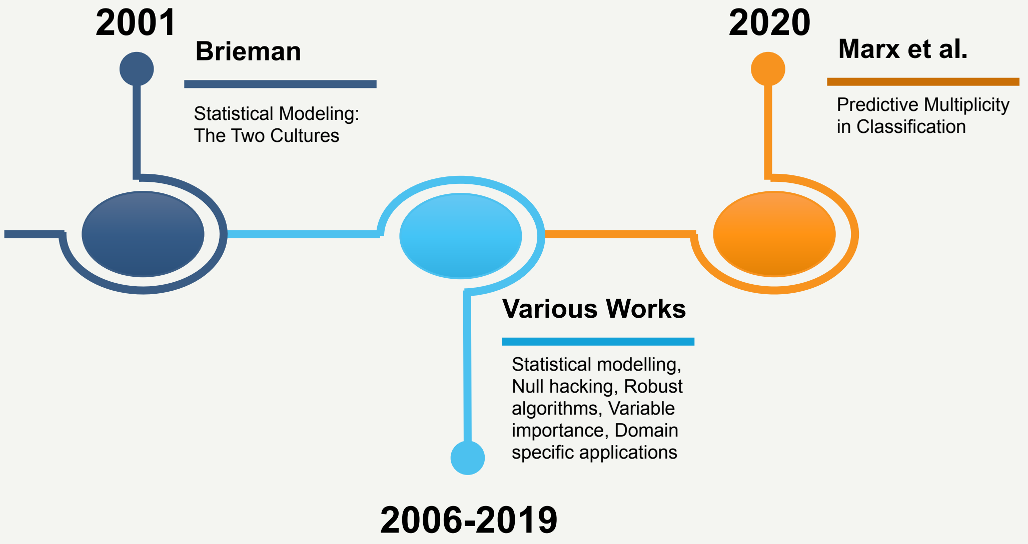
Timeline Inforgraphic from https://www.presentationgo.com

From Rashomon Effect to Multiplicity



Timeline Inforgraphic from https://www.presentationgo.com

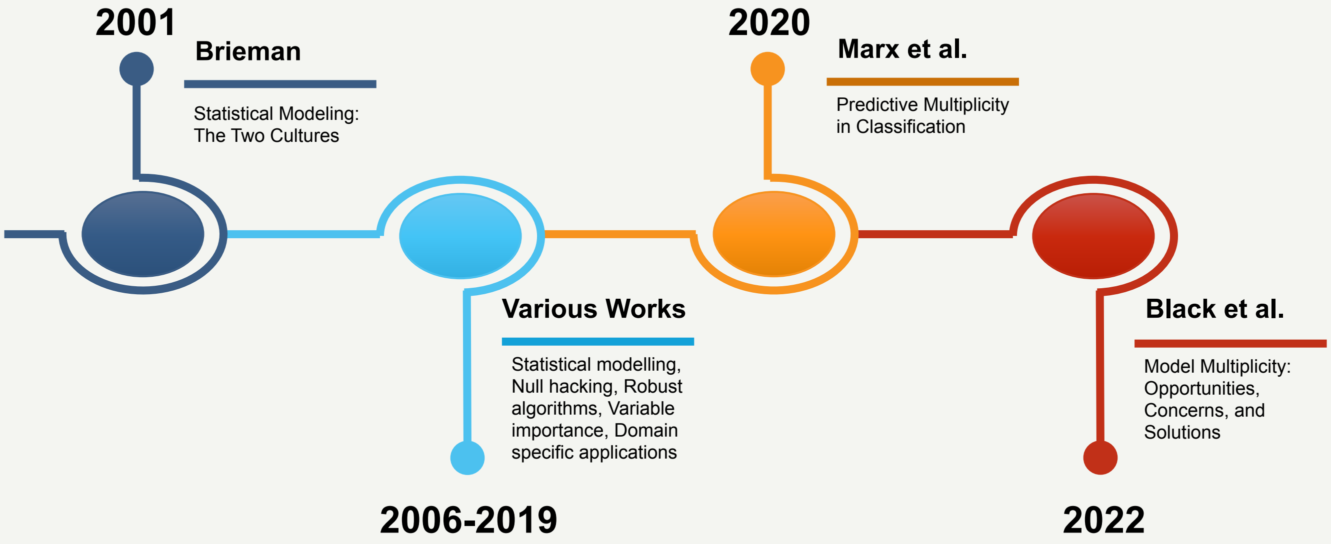
From Rashomon Effect to Multiplicity



Timeline Inforgraphic from https://www.presentationgo.com

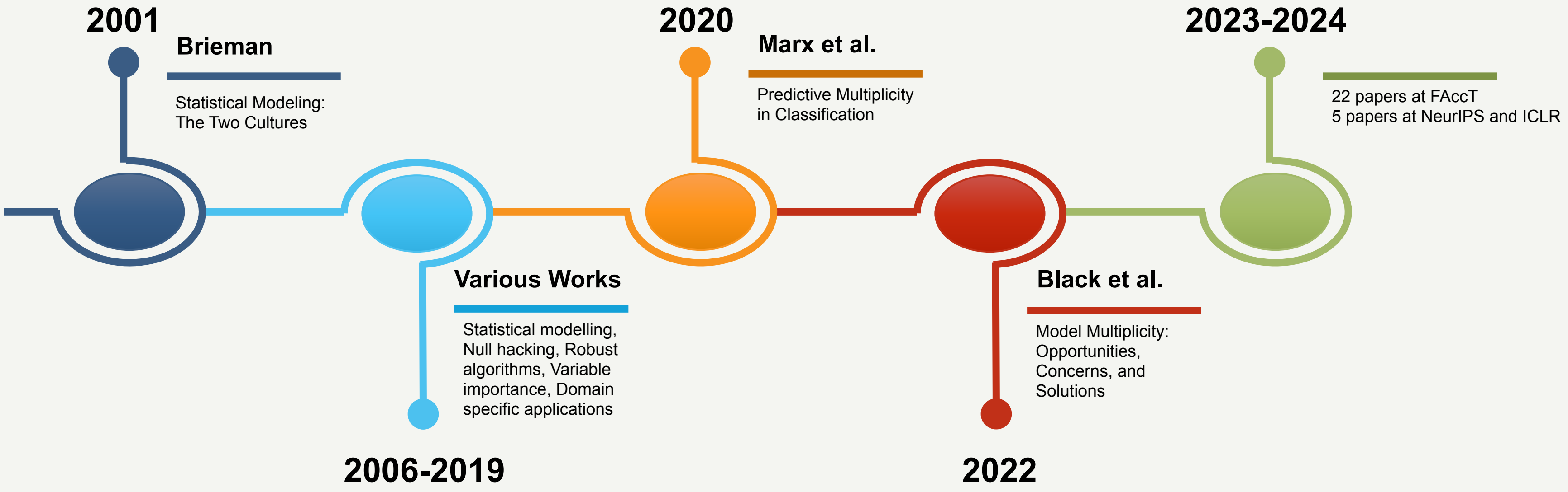
Mila

From Rashomon Effect to Multiplicity



Timeline Inforgraphic from https://www.presentationgo.com

From Rashomon Effect to Multiplicity



Timeline Inforgraphic from https://www.presentationgo.com

Model Multiplicity: Opportunities, Concerns, and Solutions

Emily Black, Manish Raghavan, and Solon Barocas

FAccT 2022

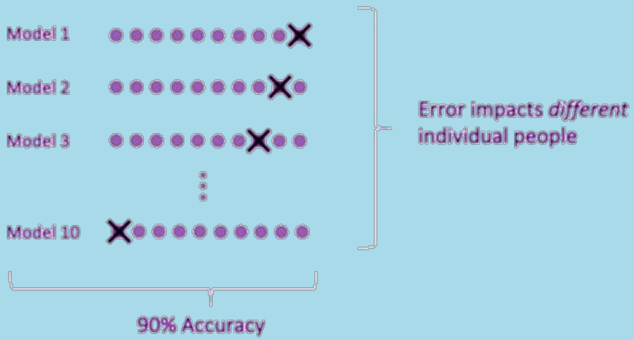


Image Credit: Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). *Less discriminatory algorithms*. Georgetown Law Journal, 113(1).

Model Multiplicity

Model Multiplicity

“Model multiplicity occurs when models with equivalent accuracy for a certain prediction task differ in terms of their internals or their predictions.”

Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Model Multiplicity

"Model multiplicity occurs when models with equivalent accuracy for a certain prediction task differ in terms of their internals or their predictions."

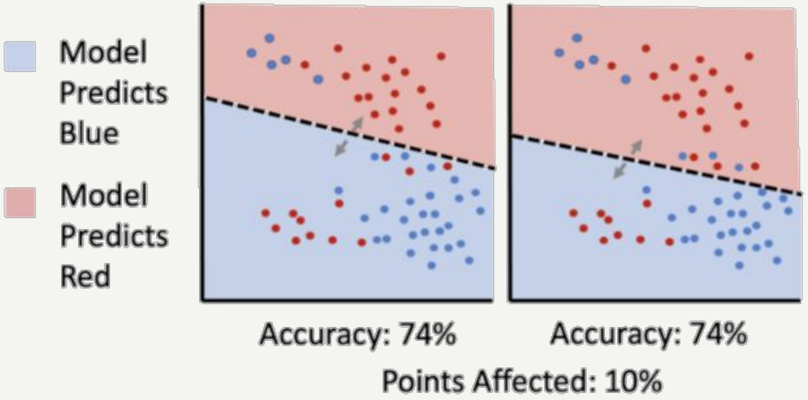


Image Credit: Black et al. 2022

Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Model Multiplicity

"Model multiplicity occurs when models with equivalent accuracy for a certain prediction task differ in terms of their internals or their predictions."

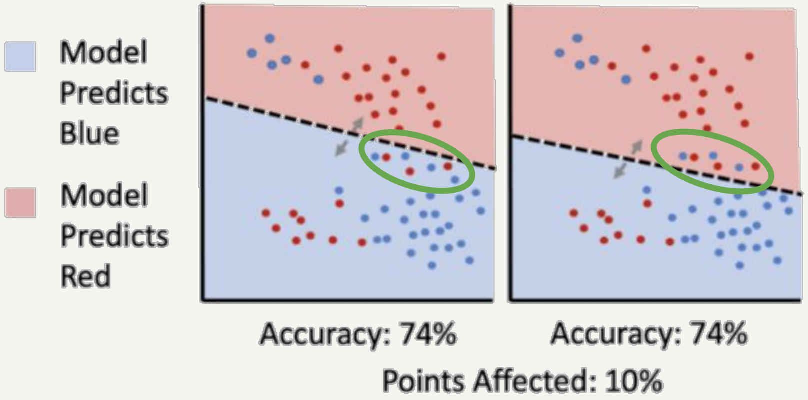


Image Credit: Black et al. 2022

Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Model Multiplicity: Opportunities

Model Multiplicity: Opportunities

Flexibility!

Model Multiplicity: Opportunities

Flexibility!

- Flexibility to choose models with better fairness (Perrone et al. 2021, Ganesh et al. 2023, 2024, Black et al. 2024).

- Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021, July). *Fair bayesian optimization*. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 854-863).
- Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).
- Ganesh, P., Gohar, U., Cheng, L., & Farnadi, G. (2024). *Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML*. Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation, NeurIPS.
- Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). *Less discriminatory algorithms*. Georgetown Law Journal, 113(1).

Model Multiplicity: Opportunities

Flexibility!

- Flexibility to choose models with better fairness (Perrone et al. 2021, Ganesh et al. 2023, 2024, Black et al. 2024).
- Flexibility to choose simpler models (Semenova et al. 2022, 2024).

- Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021, July). *Fair bayesian optimization*. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 854-863).
- Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).
- Ganesh, P., Gohar, U., Cheng, L., & Farnadi, G. (2024). *Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML*. Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation, NeurIPS.
- Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). *Less discriminatory algorithms*. Georgetown Law Journal, 113(1).
- Semenova, L., Rudin, C., & Parr, R. (2022, June). *On the existence of simpler machine learning models*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1827-1858).
- Semenova, L., Chen, H., Parr, R., & Rudin, C. (2024). *A path to simpler models starts with noise*. NeurIPS, 36.

Model Multiplicity: Opportunities

Flexibility!

- Flexibility to choose models with better fairness (Perrone et al. 2021, Ganesh et al. 2023, 2024, Black et al. 2024).
- Flexibility to choose simpler models (Semenova et al. 2022, 2024).
- Flexibility to choose models that allow actionable recourse (suggested by Black et al. 2022).

- Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021, July). *Fair bayesian optimization*. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 854-863).
- Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).
- Ganesh, P., Gohar, U., Cheng, L., & Farnadi, G. (2024). *Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML*. Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation, NeurIPS.
- Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). *Less discriminatory algorithms*. Georgetown Law Journal, 113(1).
- Semenova, L., Rudin, C., & Parr, R. (2022, June). *On the existence of simpler machine learning models*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1827-1858).
- Semenova, L., Chen, H., Parr, R., & Rudin, C. (2024). *A path to simpler models starts with noise*. NeurIPS, 36.
- Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Mila

Model Multiplicity: Opportunities

Flexibility!

- Flexibility to choose models with better fairness (Perrone et al. 2021, Ganesh et al. 2023, 2024, Black et al. 2024).
- Flexibility to choose simpler models (Semenova et al. 2022, 2024).
- Flexibility to choose models that allow actionable recourse (suggested by Black et al. 2022).
- And much more...

- Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., & Archambeau, C. (2021, July). *Fair bayesian optimization*. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 854-863).
- Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).
- Ganesh, P., Gohar, U., Cheng, L., & Farnadi, G. (2024). *Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML*. Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation, NeurIPS.
- Black, E., Koepke, J. L., Kim, P., Barocas, S., & Hsu, M. (2024). *Less discriminatory algorithms*. Georgetown Law Journal, 113(1).
- Semenova, L., Rudin, C., & Parr, R. (2022, June). *On the existence of simpler machine learning models*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1827-1858).
- Semenova, L., Chen, H., Parr, R., & Rudin, C. (2024). *A path to simpler models starts with noise*. NeurIPS, 36.
- Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Model Multiplicity: Concerns

Model Multiplicity: Concerns

Underspecification

Model Multiplicity: Concerns

Underspecification

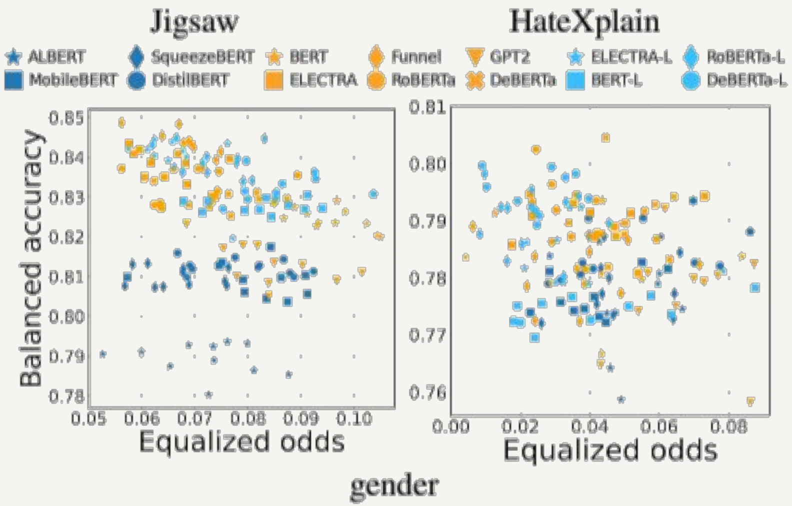


Image Credit: Baldini et al. 2022

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

Mila

Model Multiplicity: Concerns

Underspecification

Anders et al. 2020: “*We show that for any classifier g, one can always find another classifier g’ which agrees with the original g on the entire data manifold but has (almost) completely controlled explanations.*”

Anders, C., Pasliev, P., Dombrowski, A. K., Müller, K. R., & Kessel, P. (2020, November). *Fairwashing explanations with off-manifold detergent*. In International Conference on Machine Learning (pp. 314-323). PMLR.

Model Multiplicity: Concerns

Loss of Justifiability

Model Multiplicity: Concerns

Loss of Justifiability

Why am I being subject to an adversarial prediction, if there exists another model which would have given me a positive prediction, while maintaining the same overall accuracy?

Model Multiplicity: Concerns

Arbitrariness

Model Multiplicity: Concerns

Arbitrariness

Gomez et al. 2024: "… *varying the random seed can dramatically alter the predictions… This observation is equivalent to a judge sometimes flipping coins to [make a judgement].*"

Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024, June). *Algorithmic Arbitrariness in Content Moderation*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2234-2253).

Model Multiplicity: Concerns

Arbitrariness

Gomez et al. 2024: “... *varying the random seed can dramatically alter the predictions... This observation is equivalent to a judge sometimes flipping coins to [make a judgement].*”

Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024, June). *Algorithmic Arbitrariness in Content Moderation*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2234-2253).

Model Multiplicity: Demo

Let's train our own models!

https://huggingface.co/spaces/prakharg24/multiplicity-demo

Definitions, Metrics, and Nomenclature

Rashomon Sets

Or a set of competing models, a set of good models, ε-Rashomon set, ε-Level set, etc.

Rashomon Sets

Or a set of competing models, a set of good models, ε-Rashomon set, ε-Level set, etc.

Definition (ε-Level Set): Given a baseline classifier h_0 and a hypothesis class $\mathcal{H}$, the ε-level set around h_0 is the set of all classifiers $h \in \mathcal{H}$ with an error rate of at most $L(h_0) + \varepsilon$,

$$S_\varepsilon(h_0) := \{h \in \mathcal{H} : L(h) \leq L(h_0) + \varepsilon\}$$

Marx, C., Calmon, F., & Ustun, B. (2020, November). *Predictive multiplicity in classification*. In International Conference on Machine Learning (pp. 6765-6774). PMLR.

Rashomon Sets

Or a set of competing models, a set of good models, ε-Rashomon set, ε-Level set, etc.

Definition (ε-Level Set): Given ~~a baseline classifier~~ $\cancel{h_0}$ ~~and~~ a hypothesis class $\mathcal{H}$, the ε-level set ~~around~~ $\cancel{h_0}$ is the set of all classifiers $h \in \mathcal{H}$ with an error rate of at most $\cancel{L(h_0) +}$ ε,

$$S_\varepsilon(h_0) := \{h \in \mathcal{H} : L(h) \leq \varepsilon\}$$

Hsu, H., & Calmon, F. (2022). *Rashomon capacity: A metric for predictive multiplicity in classification*. Advances in Neural Information Processing Systems, 35, 28988-29000.

Predictive Multiplicity

Definition (Predictive Multiplicity): Given a baseline classifier h_0, a prediction problem exhibits predictive multiplicity over the ε-level set $S_\varepsilon(h_0)$ if there exists a model $h \in S_\varepsilon(h_0)$ such that $h(\boldsymbol{x_i}) \neq h_0(\boldsymbol{x_i})$ for some $\boldsymbol{x_i}$ in the training set.

Marx, C., Calmon, F., & Ustun, B. (2020, November). *Predictive multiplicity in classification*. In International Conference on Machine Learning (pp. 6765-6774). PMLR.

Ambiguity

Definition (Ambiguity): The ambiguity of a prediction problem over the ε-level set $S_\varepsilon(h_0)$ is the proportion of points in a training dataset that can be assigned conflicting prediction by a competing classifier $h \in S_\varepsilon(h_0)$.

Marx, C., Calmon, F., & Ustun, B. (2020, November). *Predictive multiplicity in classification*. In International Conference on Machine Learning (pp. 6765-6774). PMLR.

Self-Consistency

Definition (Self-Consistency): Self-consistency models the probability that two models from the ε-level set $S_\varepsilon(h_0)$ will agree on their predictions for the same instance $\boldsymbol{x}_i$.

Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., ... & Zhang, B. (2024, March). *Arbitrariness and social prediction: The confounding role of variance in fair classification*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22004-22012).

Multiplicity in Probabilistic Classification

Definition (Viable Prediction Range): The viable prediction range is the smallest and largest probability estimates assigned to example $\boldsymbol{x}_i$ over ε-level set $S_\varepsilon(h_0)$.

Definition (Ambiguity): The (ε, δ)-ambiguity of a probabilistic classification task is the proportion of points in a training dataset whose probability estimates changes by at least δ over ε-level set $S_\varepsilon(h_0)$.

Watson-Daniels, J., Parkes, D. C., & Ustun, B. (2023, June). *Predictive multiplicity in probabilistic classification*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 10306-10314).

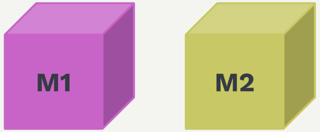
Some Other Settings

- Multiplicity for Object Detection (Hsu et al. 2024)
- Multi-Target Multiplicity (Watson-Daniels et al. 2023)
- Dataset Multiplicity (Meyer et al. 2023)

- Hsu, H., Li, G., Hu, S., & Chen, C. F. (2024). *Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation*. In The Twelfth International Conference on Learning Representations.
- Watson-Daniels, J., Barocas, S., Hofman, J. M., & Chouldechova, A. (2023, June). *Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 297-311).
- Meyer, A. P., Albarghouthi, A., & D'Antoni, L. (2023, June). *The dataset multiplicity problem: How unreliable data impacts predictions*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 193-204).

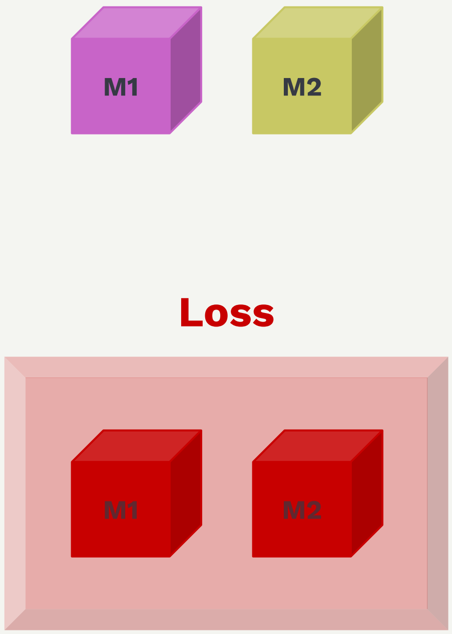
Rashomon Sets and Multiplicity

Rashomon Sets and Multiplicity

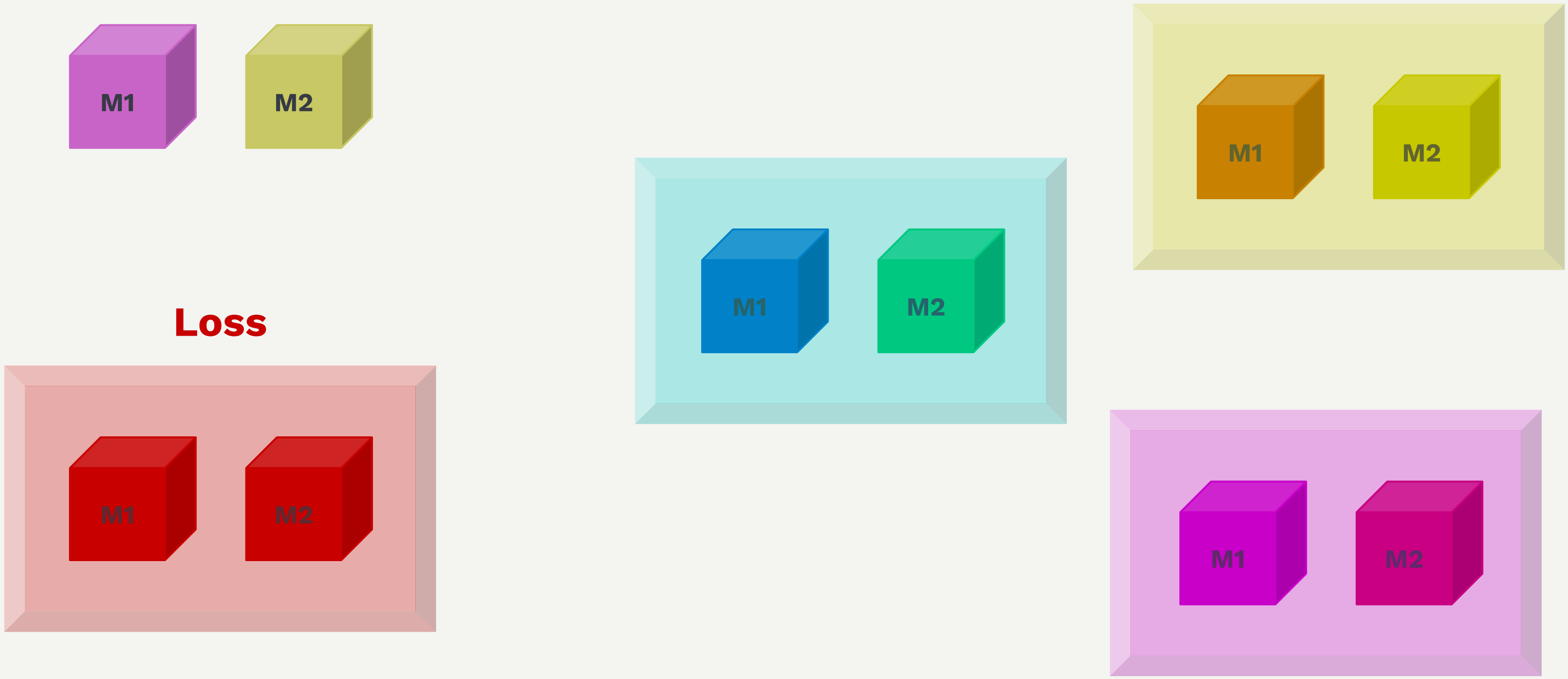


Mila

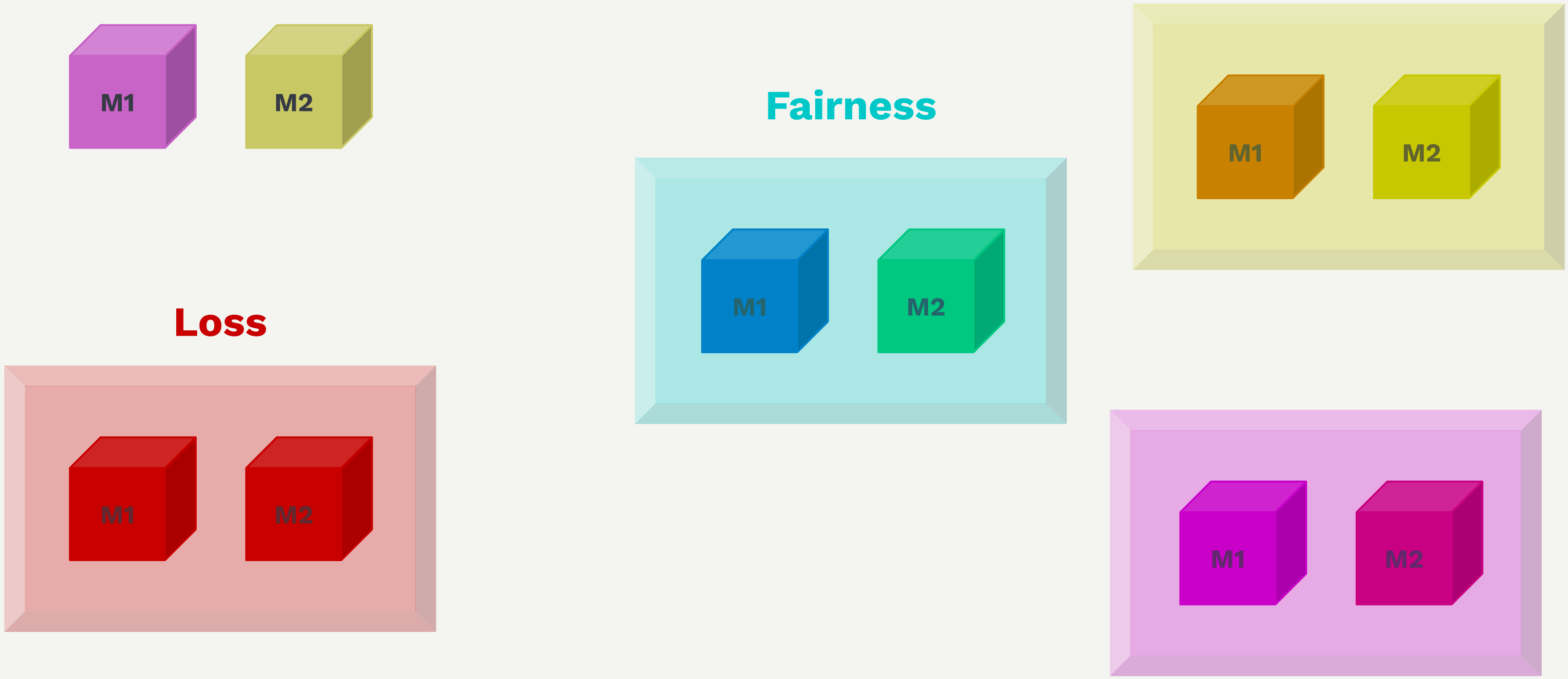
Rashomon Sets and Multiplicity



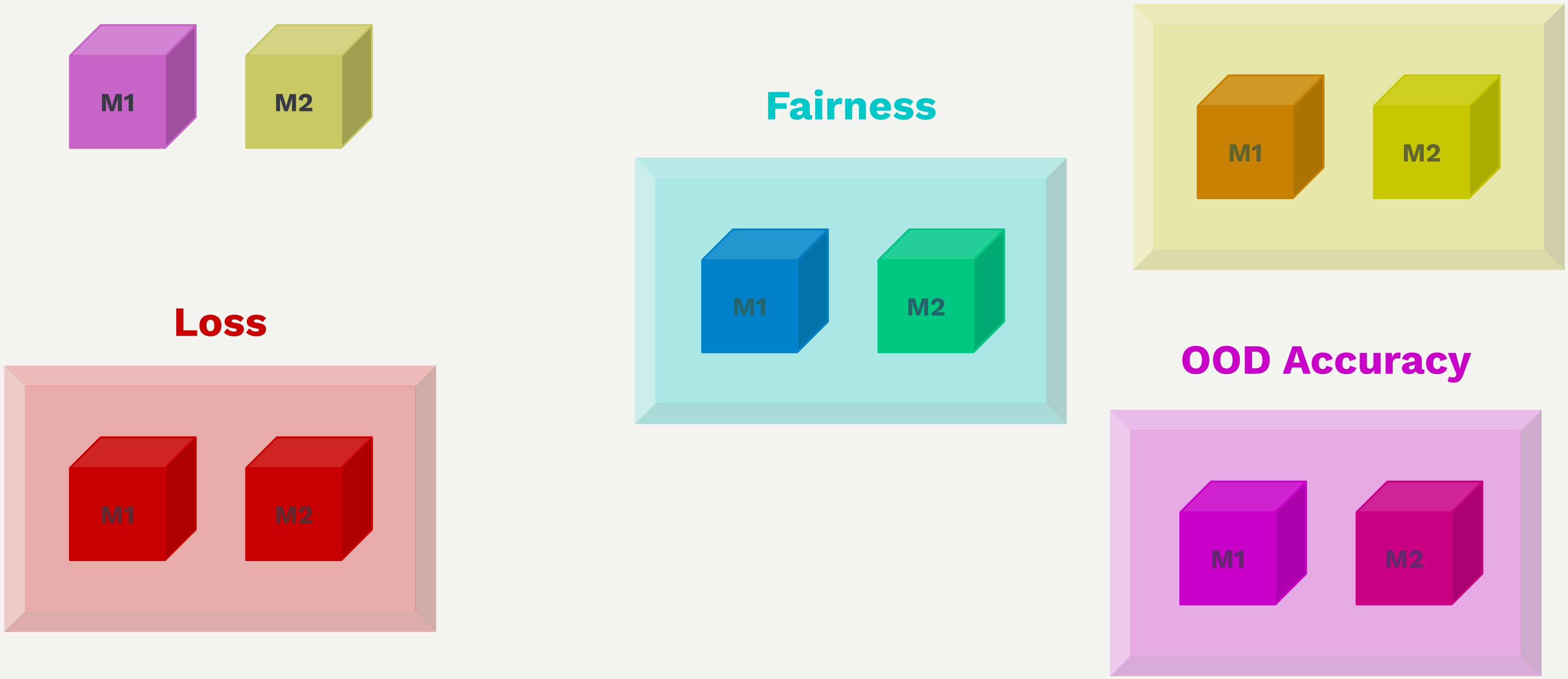
Rashomon Sets and Multiplicity



Rashomon Sets and Multiplicity



Rashomon Sets and Multiplicity



Rashomon Sets and Multiplicity

Arbitrary vs Random

Arbitrary vs Random

Arbitrary: "*... a completely unconsidered decision—one that is made without thought or perhaps even without knowledge that a choice was being made*"

Random: "*... a decision which is purposefully left to chance.*"

Black, E., Raghavan, M., & Barocas, S. (2022, June). *Model multiplicity: Opportunities, concerns, and solutions*. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 850-863).

Arbitrary vs Random

Consider two candidates for a job application with the exact same set of qualifications. But there is only one position!

One might argue that the fair thing to do is flip a coin.

Arbitrary vs Random

Consider two candidates for a job application with the exact same set of qualifications. But there is only one position!

One might argue that the fair thing to do is flip a coin.

"the decision to opt for chance must be reasoned."

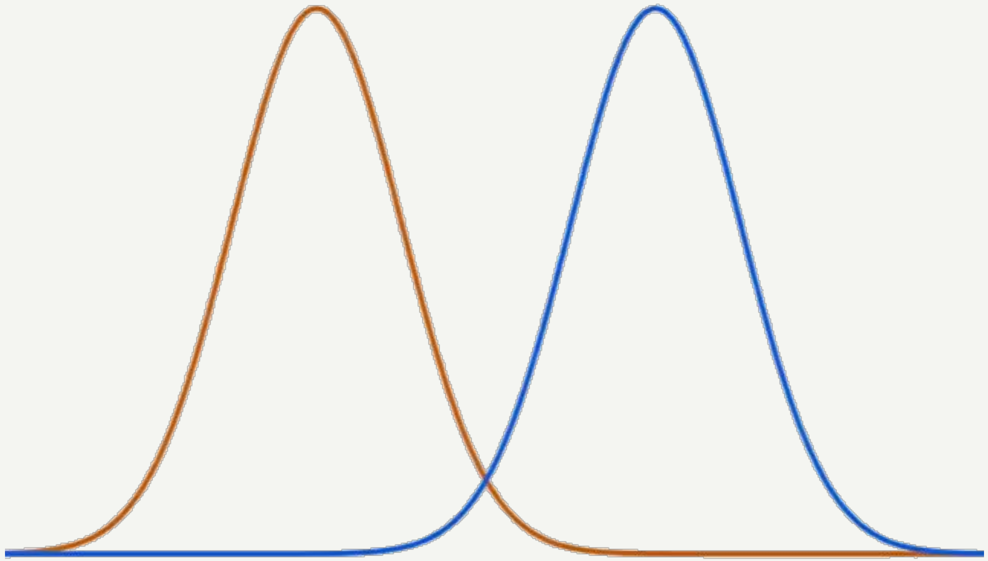
Perry, R., & Zarsky, T. (2015). *'May the Odds Be Ever in Your Favor': Lotteries in Law*. Alabama Law Review, 66, 1035-1098.

Causes of Multiplicity

https://www.educba.com/machine-learning-pipeline/

Rashomon Effect and Finite Data

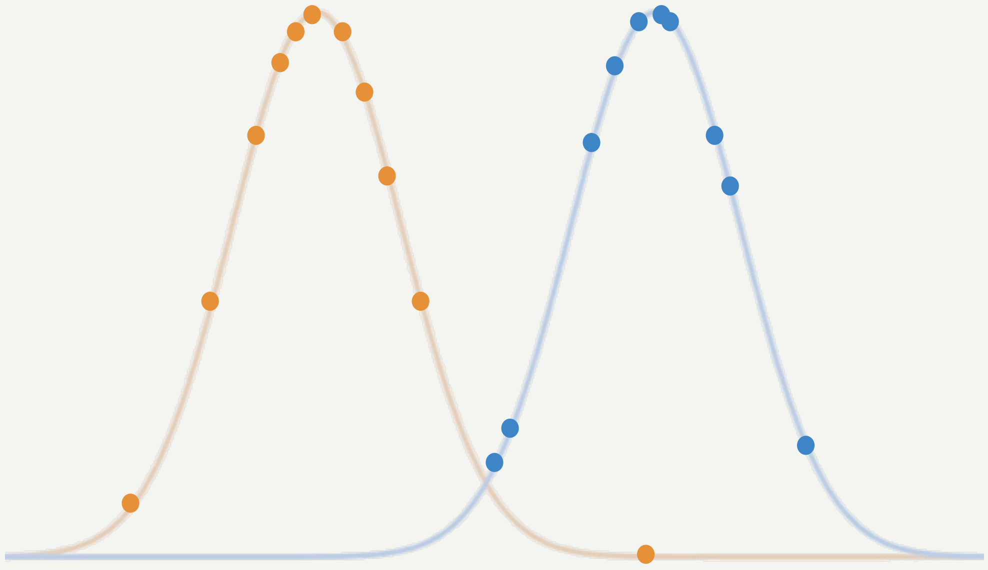
Rashomon Effect and Finite Data



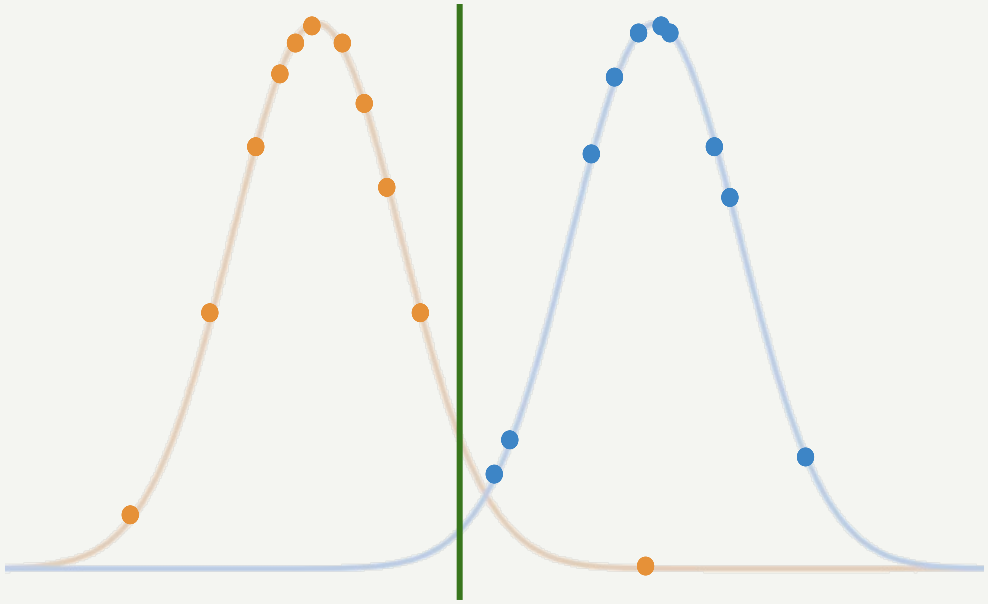
Rashomon Effect and Finite Data

The perfect classifier?

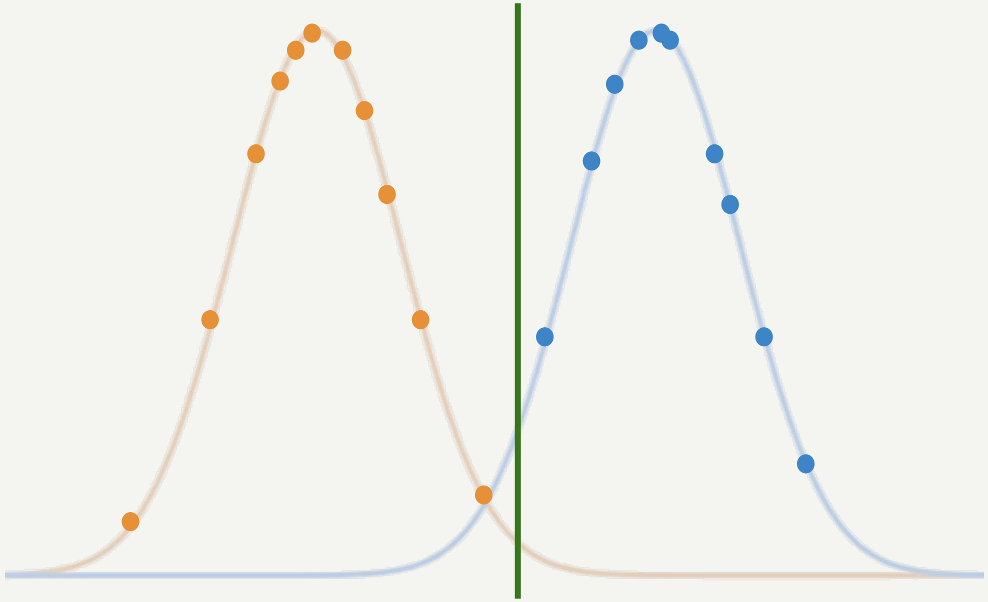
Rashomon Effect and Finite Data



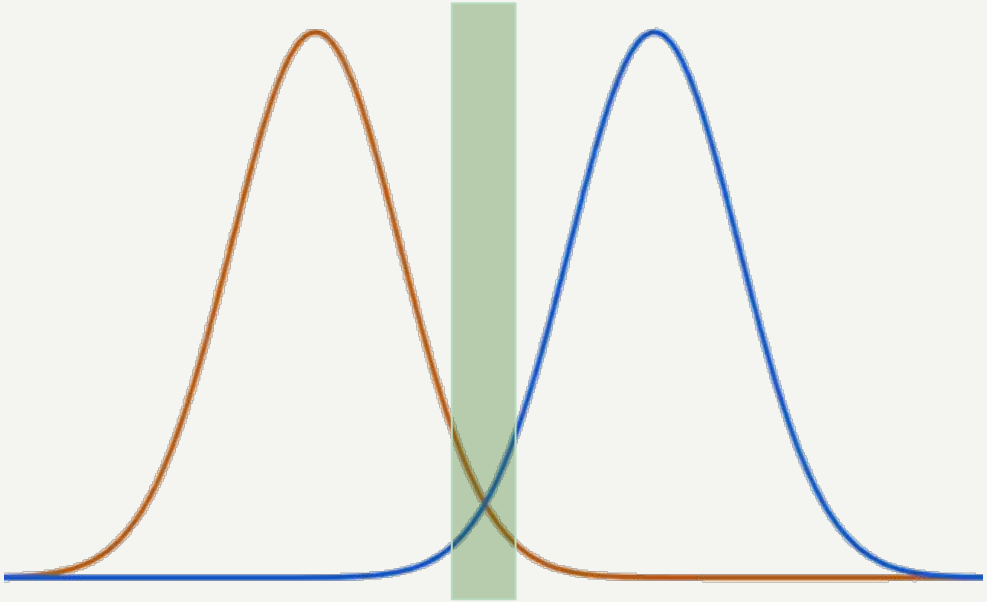
Rashomon Effect and Finite Data



Rashomon Effect and Finite Data



Rashomon Effect and Finite Data



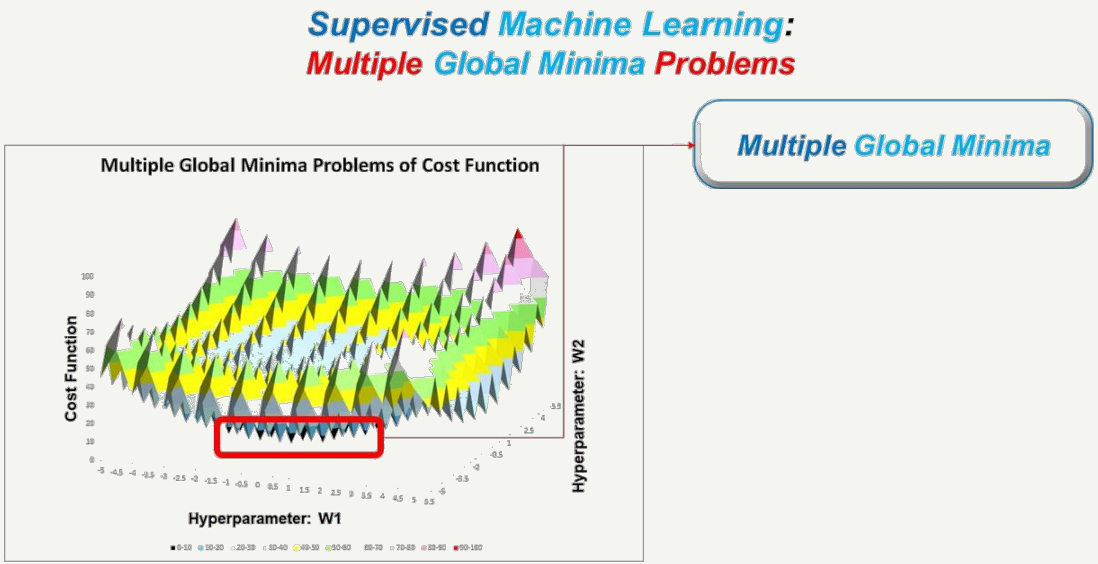
Rashomon Effect and Finite Data

“when only a finite amount of data is available to test performance, the Rashomon effect is inevitable, i.e., there exists a Rashomon set such that the empirical loss of any two models in this set are indistinguishable.”

Paes, L. M., Cruz, R., Calmon, F. P., & Diaz, M. (2023, June). *On the inevitability of the Rashomon effect*. In 2023 IEEE International Symposium on Information Theory (ISIT) (pp. 549-554). IEEE.

Rashomon Effect and Underspecification

Rashomon Effect and Underspecification



Michio Suginoo (Deep Origami), CFA: https://www.linkedin.com/in/reversalpoint/

http://www.reversalpoint.com/deep-learning-underspecification.html

Mila

Rashomon Effect and Underspecification

"We have seen that underspecification is ubiquitous in practical machine learning pipelines across many domains. Indeed, thanks to underspecification, substantively important aspects of the decisions are determined by arbitrary choices such as the random seed used for parameter initialization."

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2022). *Underspecification presents challenges for credibility in modern machine learning*. Journal of Machine Learning Research, 23(226), 1-61.

Rashomon Effect and Underspecification

“We have seen that underspecification is ubiquitous in practical machine learning pipelines across many domains. Indeed, thanks to underspecification, substantively important aspects of the decisions are determined by arbitrary choices such as the random seed used for parameter initialization.”

“modern deep learning pipelines [also] incorporate a wide variety of ‘ad hoc’ practices… These include the particular scheme used for initialization; conventions for parameterization; choice of optimization algorithm; conventions for representing data; and choices of batch size, learning rate, and other hyperparameters, all of which may interact with the infrastructure available for training and serving models.”

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., … & Sculley, D. (2022). *Underspecification presents challenges for credibility in modern machine learning*. Journal of Machine Learning Research, 23(226), 1-61.

Different Stages of the Pipeline

Different Stages of the Pipeline

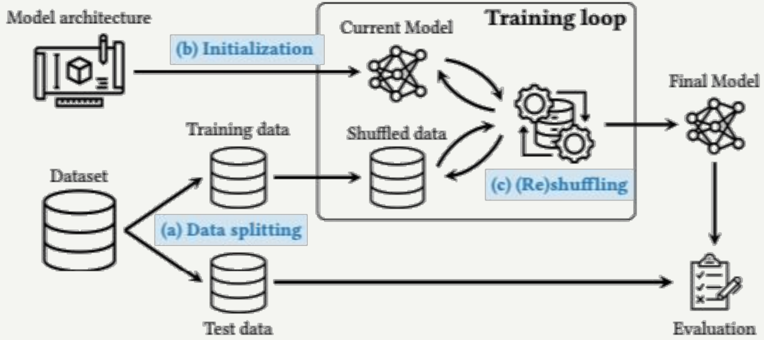


Image Credit: Ganesh et al. 2023

Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).

Different Stages of the Pipeline

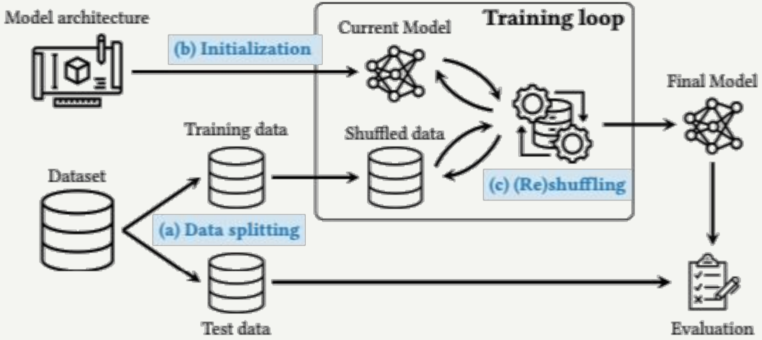
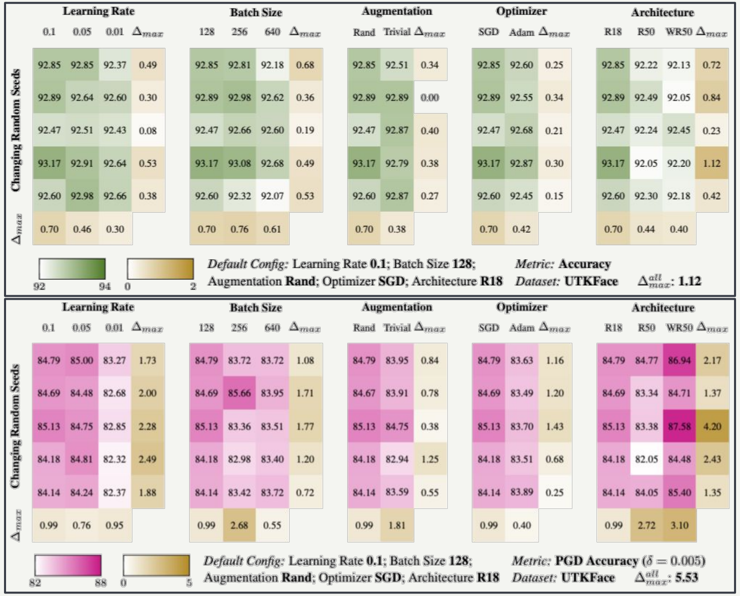


Image Credit: Ganesh et al. 2023



	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	92.85	92.85	92.37	0.49	92.85	92.81	92.18	0.68	92.85	92.51	0.34	92.85	92.60	0.25	92.85	92.22	92.13	0.72
	92.89	92.64	92.60	0.30	92.89	92.98	92.62	0.36	92.89	92.89	0.00	92.89	92.55	0.34	92.89	92.49	92.05	0.84
	92.47	92.51	92.43	0.08	92.47	92.66	92.60	0.19	92.47	92.87	0.40	92.47	92.68	0.21	92.47	92.24	92.45	0.23
	93.17	92.91	92.64	0.53	93.17	93.08	92.68	0.49	93.17	92.79	0.38	93.17	92.87	0.30	93.17	92.05	92.20	1.12
	92.60	92.98	92.66	0.38	92.60	92.32	92.07	0.53	92.60	92.87	0.27	92.60	92.45	0.15	92.60	92.30	92.18	0.42
Δ_{max}	0.70	0.46	0.30		0.70	0.76	0.61		0.70	0.38		0.70	0.42		0.70	0.44	0.40	

92 94 0 2

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18** *Metric:* **Accuracy** *Dataset:* **UTKFace** Δ^{all}_{max}: **1.12**

	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	84.79	85.00	83.27	1.73	84.79	83.72	83.72	1.08	84.79	83.95	0.84	84.79	83.63	1.16	84.79	84.77	86.94	2.17
	84.69	84.48	82.68	2.00	84.69	85.66	83.95	1.71	84.67	83.91	0.78	84.69	83.49	1.20	84.69	83.34	84.71	1.37
	85.13	84.75	82.85	2.28	85.13	83.36	83.51	1.77	85.13	84.75	0.38	85.13	83.70	1.43	85.13	83.38	87.58	4.20
	84.18	84.81	82.32	2.49	84.18	82.98	83.40	1.20	84.18	82.94	1.25	84.18	83.51	0.68	84.18	82.05	84.48	2.43
	84.14	84.24	82.37	1.88	84.14	83.42	83.72	0.72	84.14	83.59	0.55	84.14	83.89	0.25	84.14	84.05	85.40	1.35
Δ_{max}	0.99	0.76	0.95		0.99	2.68	0.55		0.99	1.81		0.99	0.40		0.99	2.72	3.10	

82 88 0 5

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18** *Metric:* **PGD Accuracy** ($\delta = 0.005$) *Dataset:* **UTKFace** Δ^{all}_{max}: **5.53**

Image Credit: Ganesh et al. 2024

Ganesh, P., Chang, H., Strobel, M., & Shokri, R. (2023, June). *On the impact of machine learning randomness on group fairness*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1789-1800).
Ganesh, P. (2024). An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4488-4497).

Different Stages of the Pipeline

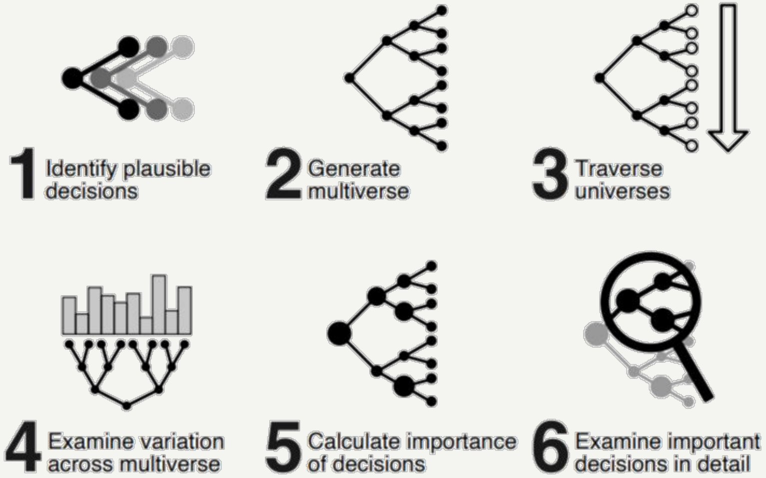


Image Credit: Simson et al. 2024

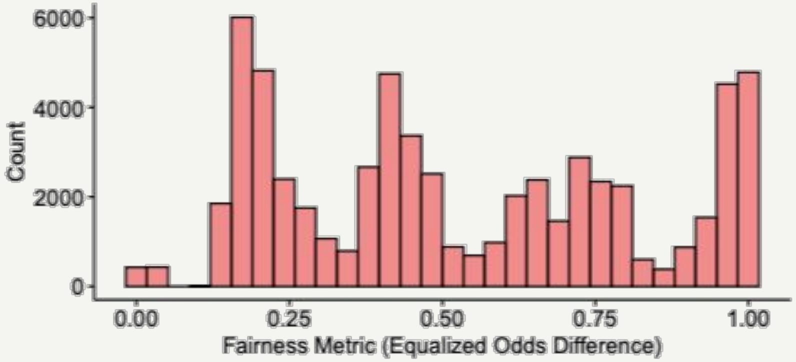


Image Credit: Simson et al. 2024

Simson, J., Pfisterer, F., & Kern, C. (2024, June). *One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1305-1320).

Different Stages of the Pipeline

"... the issue of discrimination-hacking (d-hacking) in AI systems, where the brittleness of fairness measurement and mitigation can be exploited, intentionally or unintentionally, to comply with responsible AI regulation while still deploying biased systems."

Black, E., Gillis, T., & Hall, Z. Y. (2024, June). *D-hacking*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 602-615).

Multiplicity and Responsible AI

Multiplicity and Fairness

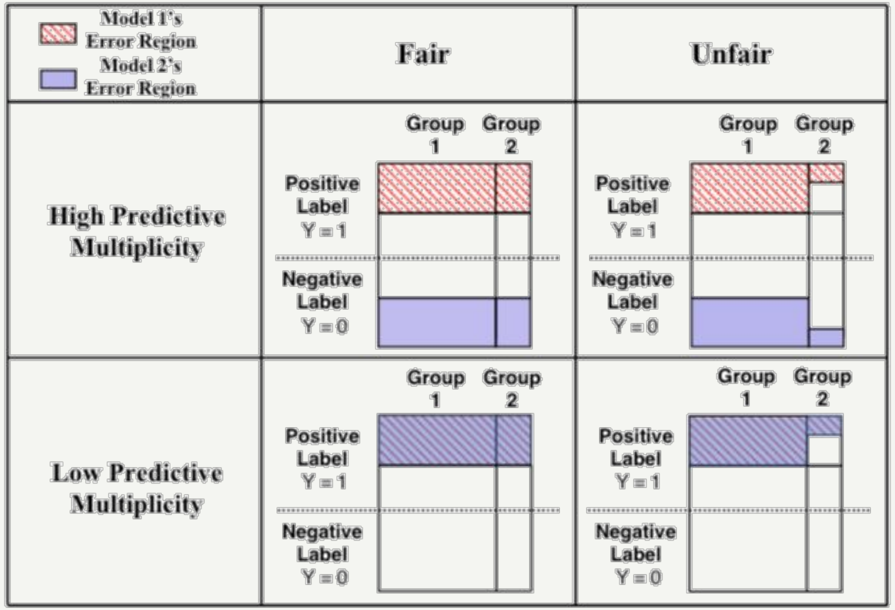


Image Credit: Long et al. 2024

Long, C., Hsu, H., Alghamdi, W., & Calmon, F. (2024). *Individual arbitrariness and group fairness*. Advances in Neural Information Processing Systems, 36.

Mila

Multiplicity and Fairness

"… fairness interventions in machine learning optimized solely for group fairness and accuracy can exacerbate predictive multiplicity. Consequently, state-of-the-art fairness interventions can mask high predictive multiplicity behind favorable group fairness and accuracy metrics."

Long, C., Hsu, H., Alghamdi, W., & Calmon, F. (2024). *Individual arbitrariness and group fairness*. Advances in Neural Information Processing Systems, 36.

Multiplicity and Privacy

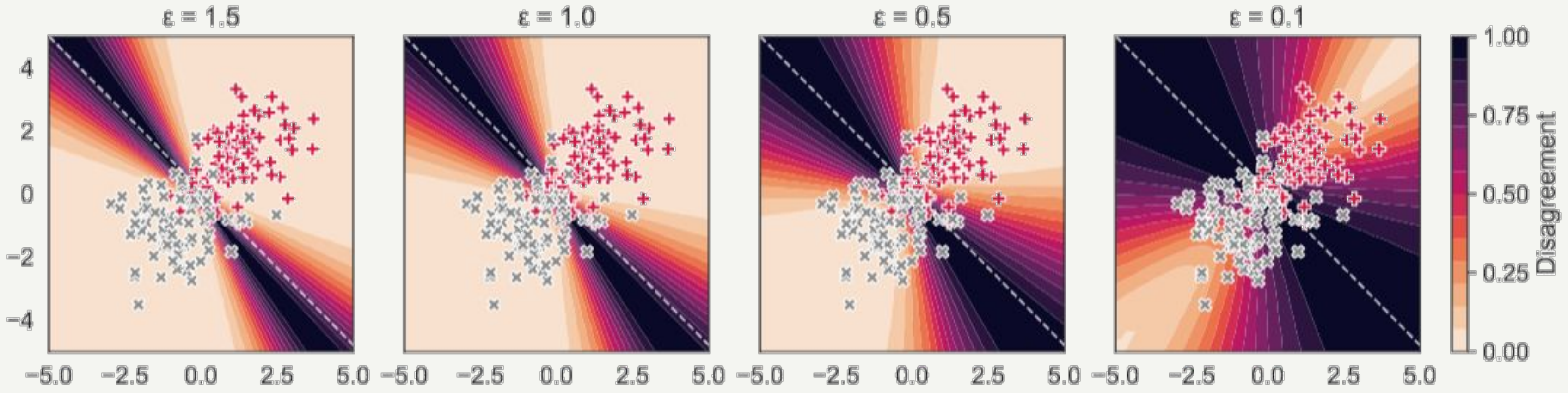


Image Credit: Kulynych et al. 2023

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

Disparity in Multiplicity

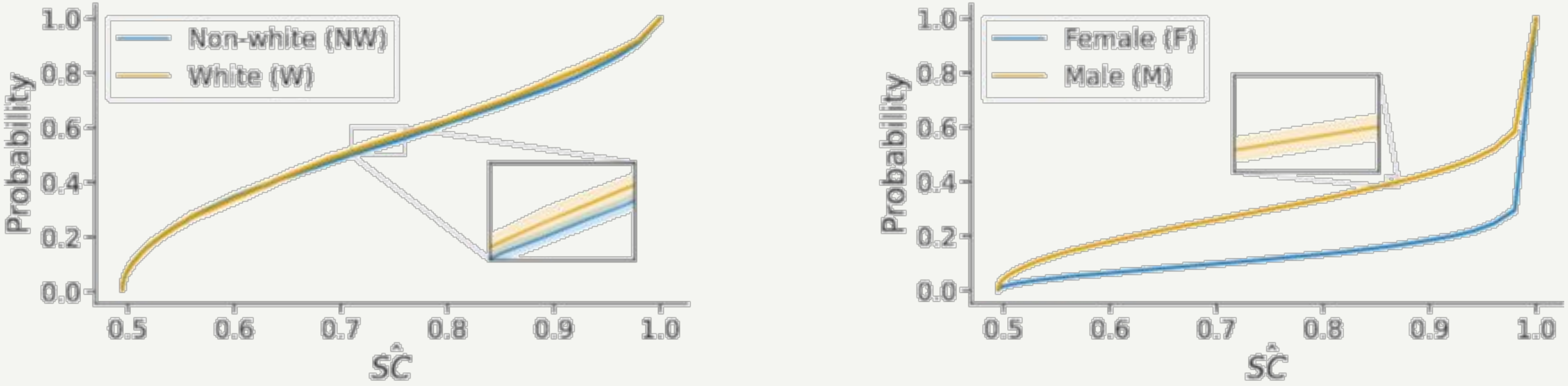


Image Credit: Cooper et al. 2024

Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., ... & Zhang, B. (2024, March). *Arbitrariness and social prediction: The confounding role of variance in fair classification*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22004-22012).

Disparity in Multiplicity

"... we show that arbitrary decisions are not uniformly spread across all texts and that it is more common in texts that target specific demographic groups (e.g., anti-LGTBQ posts)."

"We argue that predictive multiplicity poses a selective break from a rule-based approach to content moderation and infringes upon procedural fairness. [Furthermore], the [disparity] in arbitrary decisions can be discriminatory."

Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024, June). *Algorithmic Arbitrariness in Content Moderation*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2234-2253).

Disparity in Multiplicity

Under the Canadian anti-discrimination doctrine, discrimination can be defined as *“a distinction, whether intentional or not but based on grounds relating to personal characteristics of the individual or group, which has the effect of imposing burdens, obligations or disadvantages on such individual or group not imposed upon others, or which withholds or limits access to opportunities, benefits and advantages available to other members of society”*

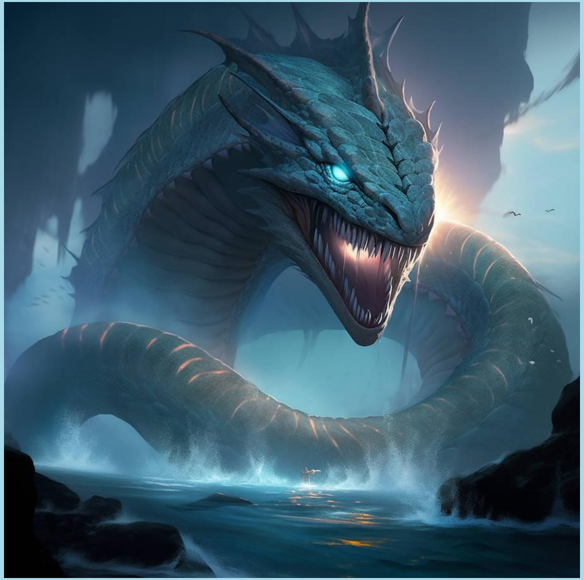
Disclaimer: The information on these slides does not, and is not intended to, constitute legal advice.

Ganesh, P., Daldaban, I., Cofone, I., & Farnadi, G. *Ongoing work.*

The Algorithmic Leviathan

Kathleen Creel and Deborah Hellman

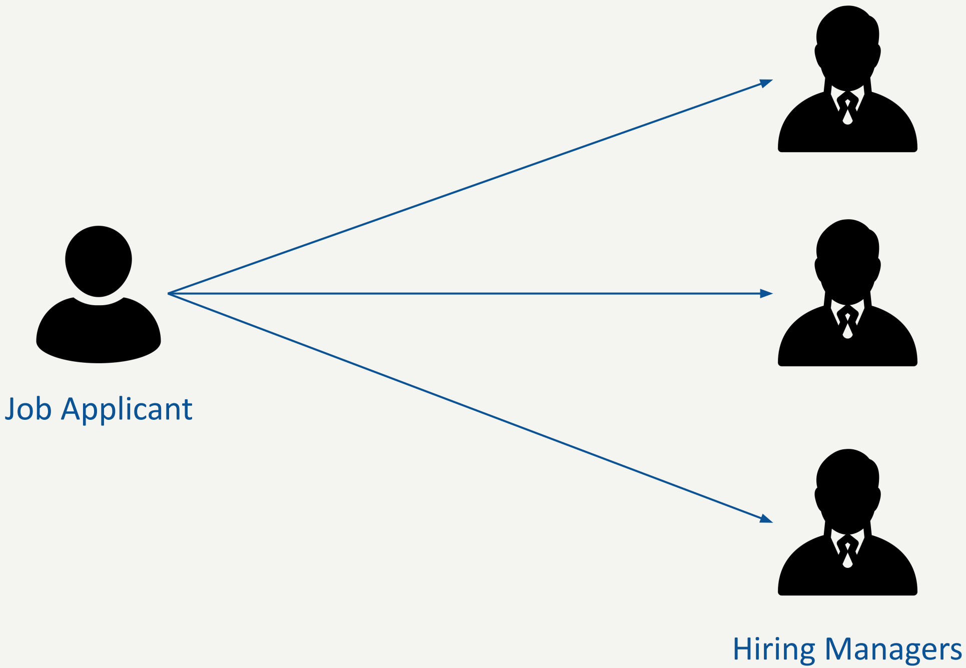
Canadian Journal of Philosophy (2022)



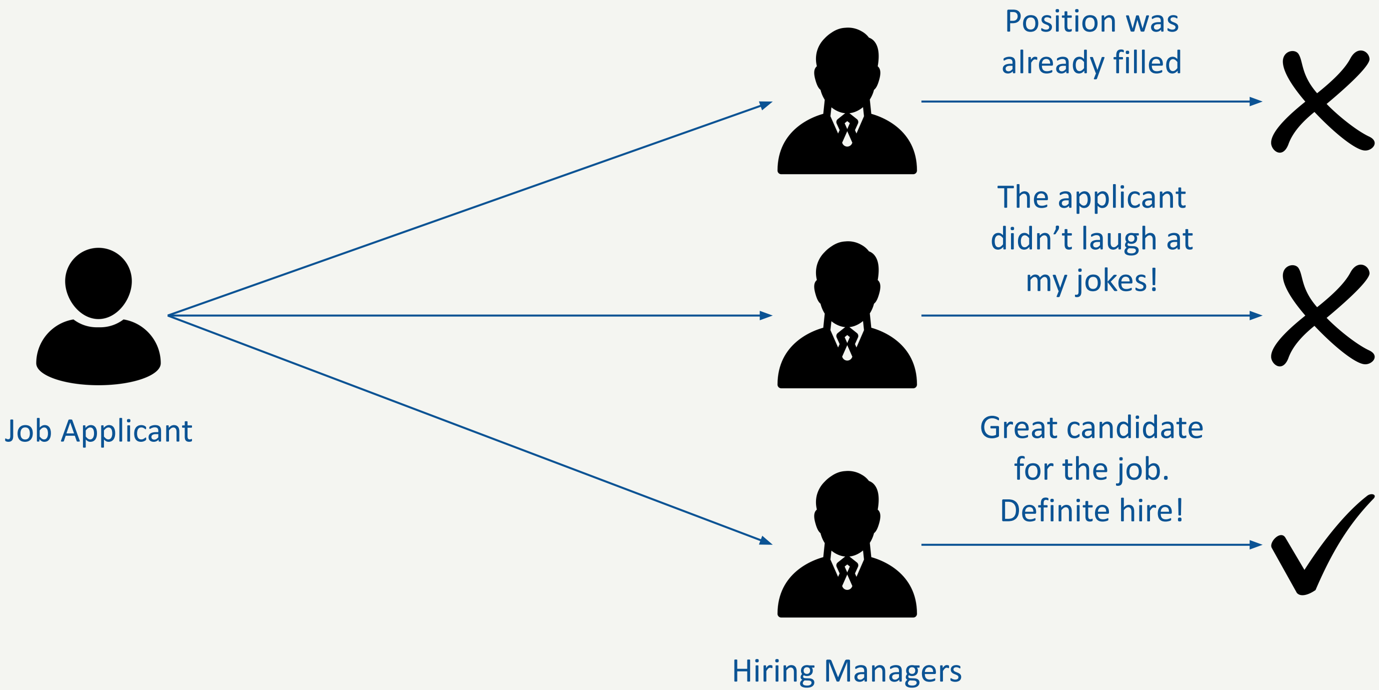
https://www.deviantart.com/alyskan/art/Leviathan-978914687

Isolated Arbitrary Decisions

Isolated Arbitrary Decisions



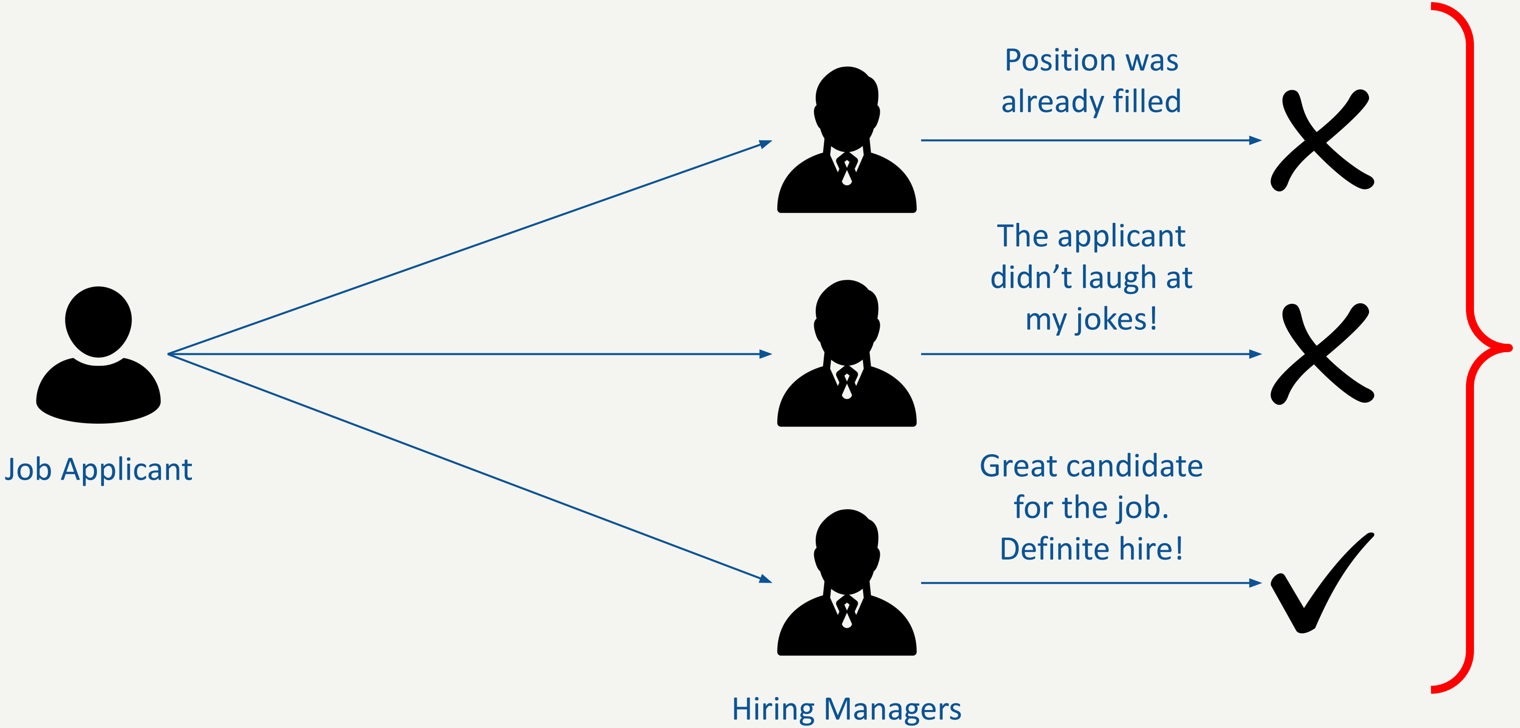
Isolated Arbitrary Decisions



Mila

Isolated Arbitrary Decisions

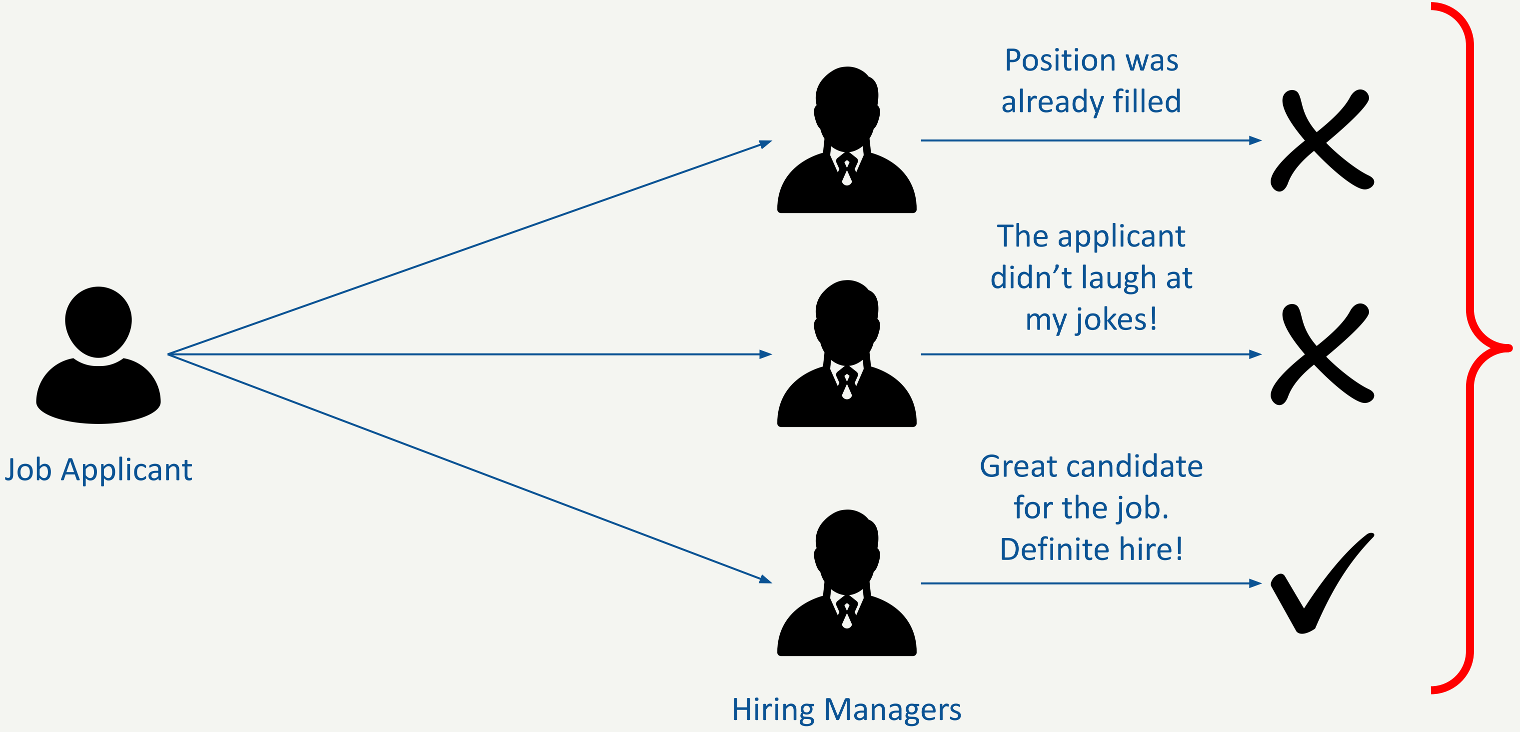
Is this arbitrariness a concern?



Isolated Arbitrary Decisions

Is this arbitrariness a concern?

Creel and Hellman argue **No**



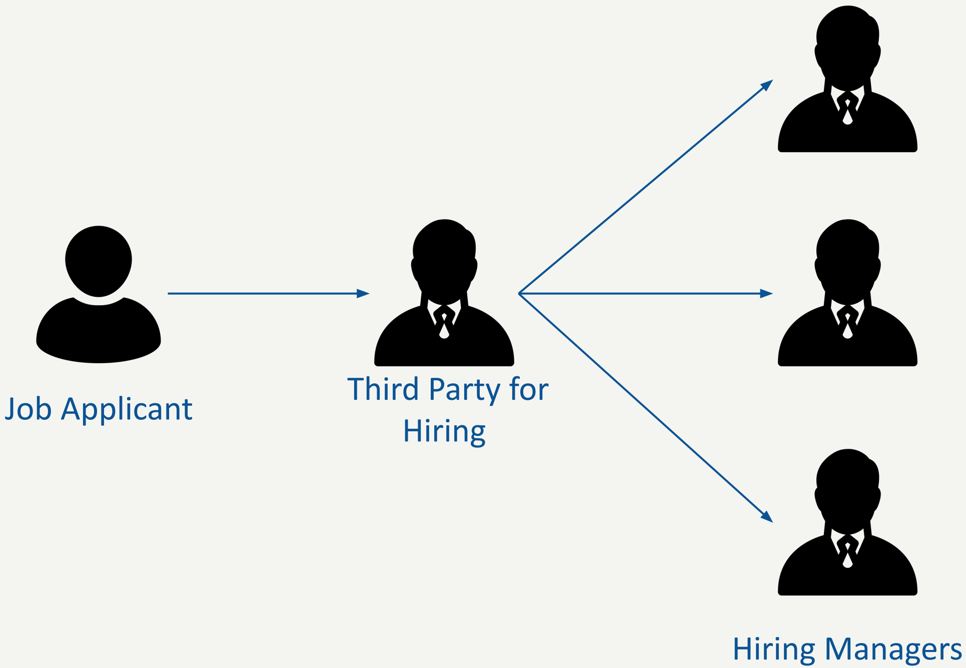
Mila

Isolated Arbitrary Decisions

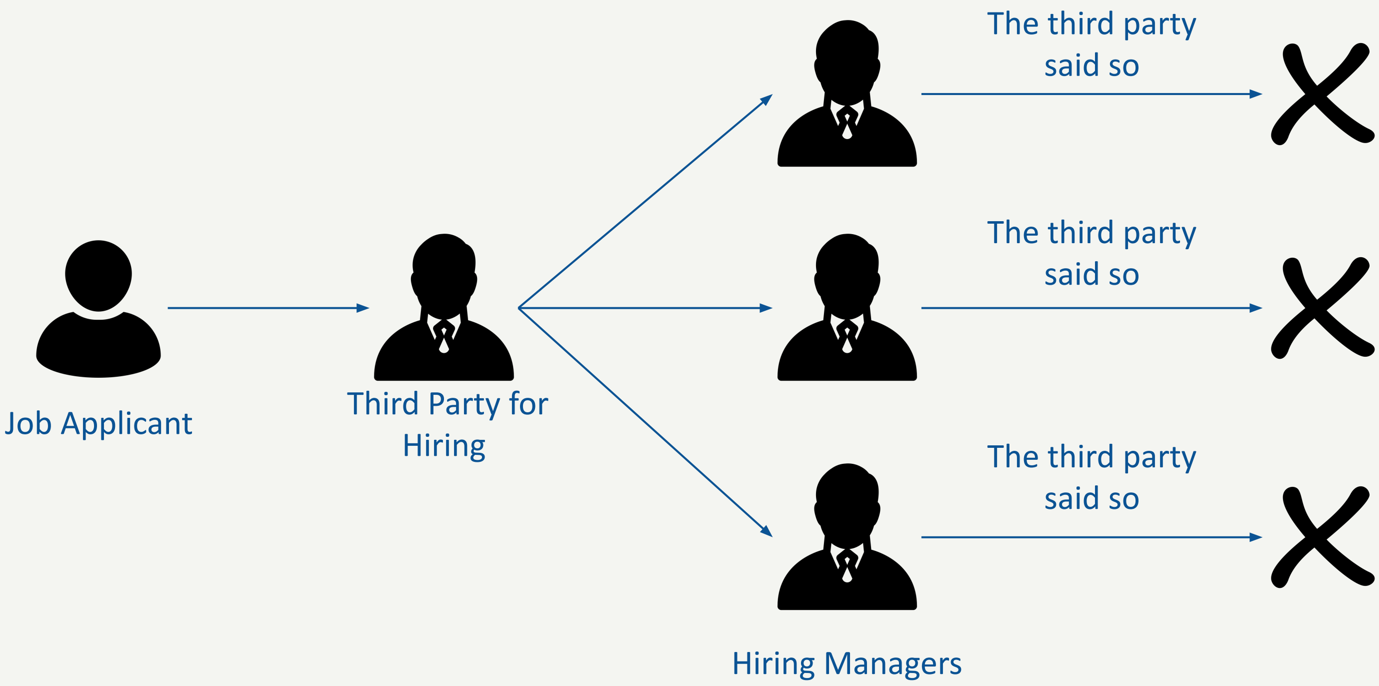
“We argue, perhaps counterintuitively, that isolated arbitrary decisions are not of moral concern except when other rights make non-arbitrariness relevant.”

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

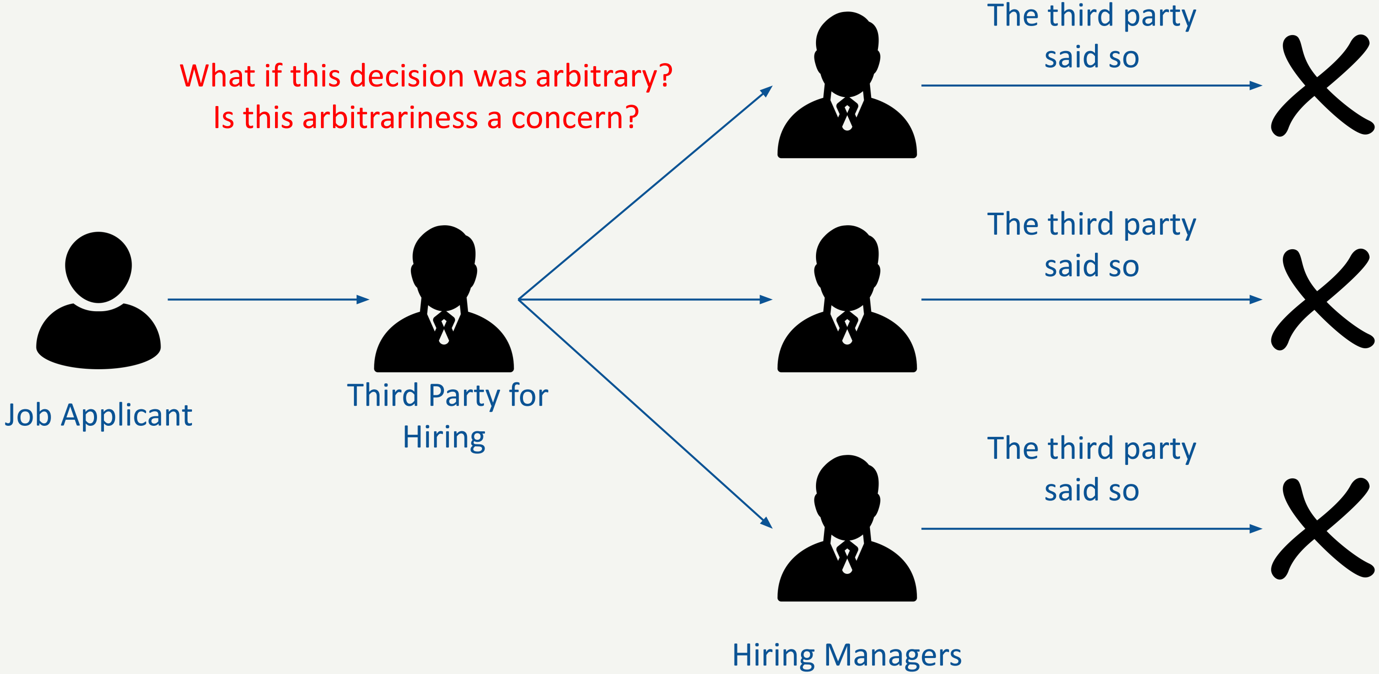
Arbitrariness at Scale



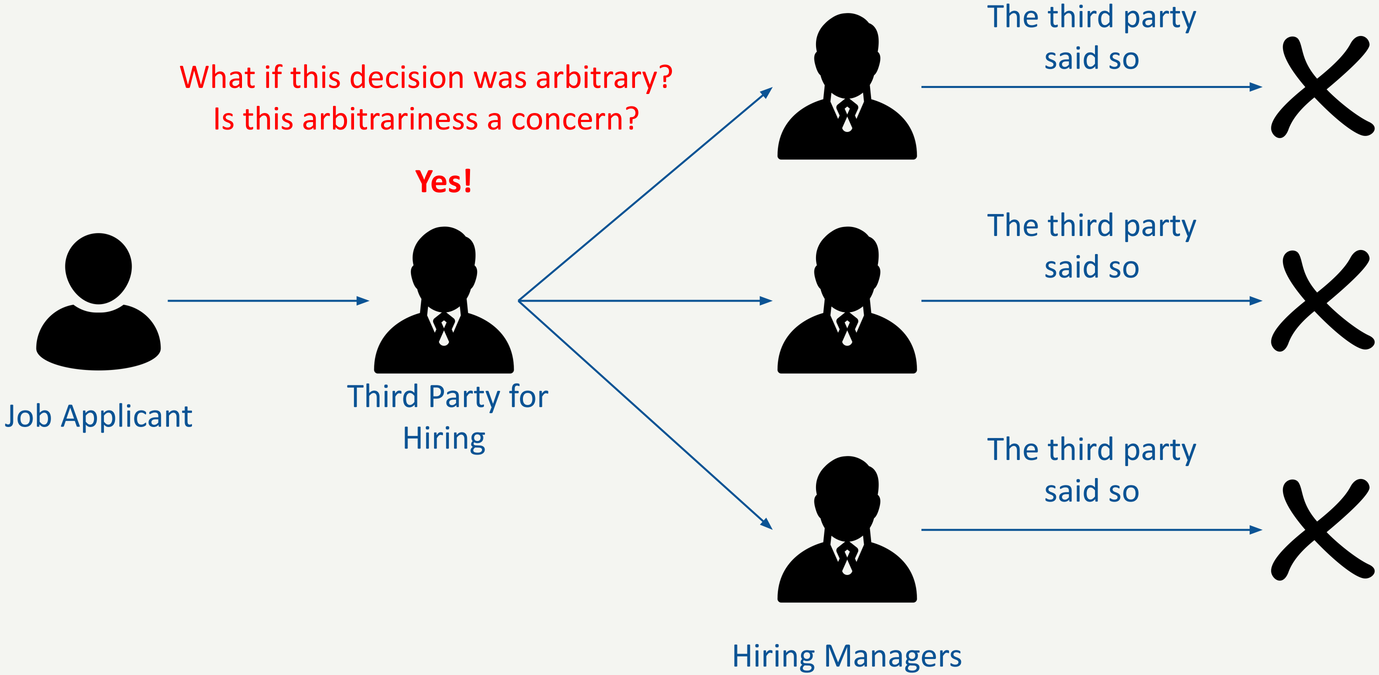
Arbitrariness at Scale



Arbitrariness at Scale



Arbitrariness at Scale



Arbitrariness at Scale

“what worries us is that the systematicity of these flawed tools means that a person affected in one context will likely be affected in multiple contexts. Because the tool is used in multiple settings, the exclusion will apply across multiple actors—lenders, for example—such that a person negatively affected will be unable to find respite with another lender.”

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

Model Multiplicity: ‘Solutions’

Auditing and Reporting Multiplicity

Auditing and Reporting Multiplicity

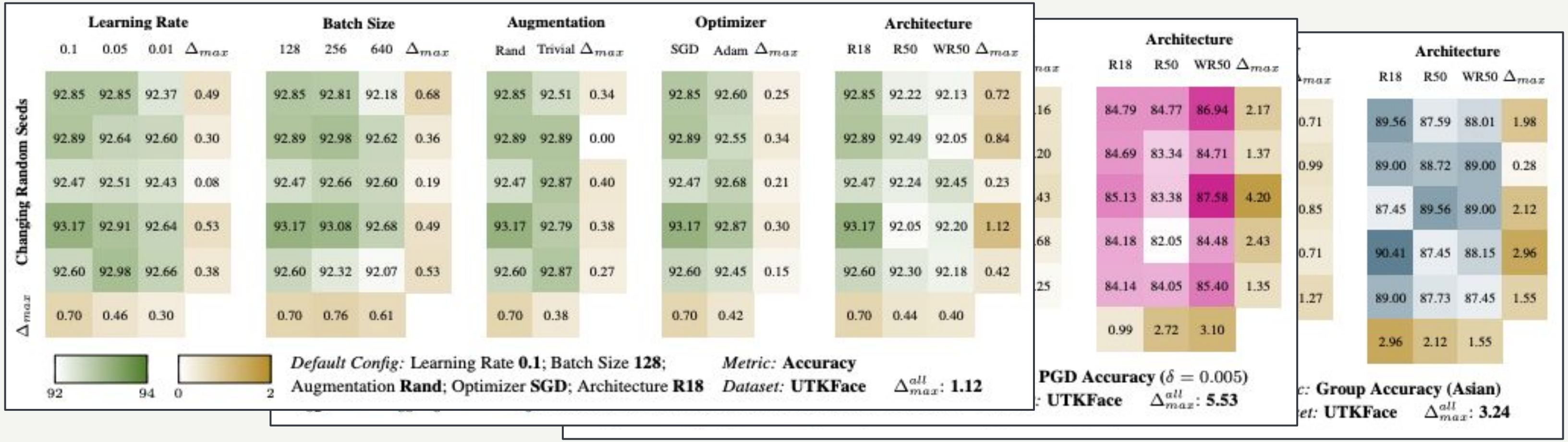
1. Create an *empirical Rashomon set*.
2. Measure and report multiplicity.

Auditing and Reporting Multiplicity

- Adversarial Weight Perturbation (AWP): To efficiently find the ‘spread’ of multiplicity.
- Dropout-Based Rashomon Set Exploration

Hsu, H., & 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., & Chen, C. F. (2024). *Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation*. In The Twelfth International Conference on Learning Representations.

Auditing and Reporting Multiplicity



Changing Random Seeds	Learning Rate 0.1	0.05	0.01	Δ_{max}	Batch Size 128	256	640	Δ_{max}	Augmentation Rand	Trivial	Δ_{max}	Optimizer SGD	Adam	Δ_{max}	Architecture R18	R50	WR50	Δ_{max}
	92.85	92.85	92.37	0.49	92.85	92.81	92.18	0.68	92.85	92.51	0.34	92.85	92.60	0.25	92.85	92.22	92.13	0.72
	92.89	92.64	92.60	0.30	92.89	92.98	92.62	0.36	92.89	92.89	0.00	92.89	92.55	0.34	92.89	92.49	92.05	0.84
	92.47	92.51	92.43	0.08	92.47	92.66	92.60	0.19	92.47	92.87	0.40	92.47	92.68	0.21	92.47	92.24	92.45	0.23
	93.17	92.91	92.64	0.53	93.17	93.08	92.68	0.49	93.17	92.79	0.38	93.17	92.87	0.30	93.17	92.05	92.20	1.12
	92.60	92.98	92.66	0.38	92.60	92.32	92.07	0.53	92.60	92.87	0.27	92.60	92.45	0.15	92.60	92.30	92.18	0.42
Δ_{max}	0.70	0.46	0.30		0.70	0.76	0.61		0.70	0.38		0.70	0.42		0.70	0.44	0.40	

Architecture R18	R50	WR50	Δ_{max}
84.79	84.77	86.94	2.17
84.69	83.34	84.71	1.37
85.13	83.38	87.58	4.20
84.18	82.05	84.48	2.43
84.14	84.05	85.40	1.35
0.99	2.72	3.10	

Architecture R18	R50	WR50	Δ_{max}
89.56	87.59	88.01	1.98
89.00	88.72	89.00	0.28
87.45	89.56	89.00	2.12
90.41	87.45	88.15	2.96
89.00	87.73	87.45	1.55
2.96	2.12	1.55	

Image Credit: Ganesh et al. 2024

Ganesh, P. (2024). An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4488-4497).

What to do after measuring multiplicity?

- Abstaining from making a decision. (Cooper et al. 2024)
- Taking a step back and re-evaluating the use of data-driven automated decision making! (Gomez et al. 2024, Neophytou et al. 2024)

Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., ... & Zhang, B. (2024, March). *Arbitrariness and social prediction: The confounding role of variance in fair classification*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22004-22012).
Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F. (2024, June). *Algorithmic Arbitrariness in Content Moderation*. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2234-2253).
Neophytou, N., Ganesh, P., Taïk, A., & Farnadi, G., (2024), *Shifting Sands: Arbitrariness in Mortgage Lending Under Distribution Shifts*. Under review

TL;DR

Model multiplicity is the phenomenon of indistinguishable models during development exhibiting diverse behavior when deployed.

Model multiplicity is the phenomenon of indistinguishable models during development exhibiting diverse behavior when deployed.

This indistinguishability creates an *arbitrariness* in selecting one model over another, with statistical, social, moral, and legal implications.

Model multiplicity is the phenomenon of indistinguishable models during development exhibiting diverse behavior when deployed.

This indistinguishability creates an *arbitrariness* in selecting one model over another, with statistical, social, moral, and legal implications.

Multiplicity is a young field with a high interdisciplinary potential!

Model multiplicity is the phenomenon of indistinguishable models during development exhibiting diverse behavior when deployed.

This indistinguishability creates an *arbitrariness* in selecting one model over another, with statistical, social, moral, and legal implications.

Multiplicity is a young field with a high interdisciplinary potential!

Interested in multiplicity?

Shoot me an email - prakhar.ganesh@mila.quebec

Slides here →