Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Systemizing Multiplicity: The Curious Case of Arbitrariness in Machine Learning

Prakhar Ganesh[1, 2], **Afaf Taïk**[3], **Golnoosh Farnadi**[1, 2]

[1]McGill University
[2]Mila - Quebec Artificial Intelligence Institute
[3] Université de Sherbrooke
prakhar.ganesh@mila.quebec, afaf.taik@usherbrooke.ca, farnadig@mila.quebec

Abstract

Algorithmic modeling relies on limited information in data to extrapolate outcomes for unseen scenarios, often embedding an element of arbitrariness in its decisions. A perspective on this arbitrariness that has recently gained interest is multiplicity—the study of arbitrariness across a set of "good models", i.e., those likely to be deployed in practice. In this work, we systemize the literature on multiplicity by: (a) formalizing the terminology around model design choices and their contribution to arbitrariness, (b) expanding the definition of multiplicity to incorporate underrepresented forms beyond just predictions and explanations, (c) clarifying the distinction between multiplicity and other lenses of arbitrariness, i.e., uncertainty and variance, and (d) distilling the benefits and potential risks of multiplicity into overarching trends, situating it within the broader landscape of responsible AI. We conclude by identifying open research questions and highlighting emerging trends in this young but rapidly growing area of research.

1 Introduction

Machine learning attempts to approximate the complexities of the world, inevitably simplifying aspects of reality and failing to fully capture its nuances (Hooker 2021; Buolamwini and Gebru 2018; Birhane 2022). It is thus inherently susceptible to arbitrariness, as it attempts to extrapolate outcomes based on limited information. Whether due to imperfect data (Buolamwini and Gebru 2018; Geirhos et al. 2020), flawed modeling assumptions (Hooker 2021; Breiman 2001; Jacobs and Wallach 2021), or inherent unknowability (Wang et al. 2024; Dressel and Farid 2018), arbitrariness is an unavoidable byproduct of data-driven learning. Hence, recognizing and understanding this arbitrariness is crucial for developing responsible learning models.

The study of arbitrariness is not new; it has long been a subject of interest in uncertainty literature, with roots going back centuries in statistics and decision theory (Savage 1972; Bayes 1763; Box and Tiao 1973; Gal 2016). Recently, however, a new paradigm called *multiplicity* has emerged. First articulated by Breiman (2001), multiplicity has gained popularity due to its focus only on the arbitrariness present within a set of "good models", i.e., models that pass certain selection criteria and thus are likely to be deployed, known as the Rashomon set. Multiplicity takes an intriguing perspective on arbitrariness in model decisions by instead examining arbitrariness in model selection. Through choices made during development, multiplicity offers an operational lens to the issue of arbitrariness and lays the groundwork for practical solutions in real-world applications.

Several works in the literature have provided broad overviews of the field of multiplicity. Black, Raghavan, and Barocas (2022) holds a special place in modern multiplicity literature, offering a discussion of "opportunities", "concerns", and potential "solutions" of multiplicity. Similarly, Rudin et al. (2024) discusses several benefits of multiplicity, with a focus on identifying simpler and more interpretable models. At this point, it would be remiss not to acknowledge the dissertations of Semenova (2024); Zhong (2024); Cooper (2024); Hsu (2023); Black (2022); Watson-Daniels (2024); Hasan (2022), contributing valuable perspectives on the role of multiplicity in machine learning. Despite these contributions, the field lacks a systematic review of the literature–clearly needed given its rapid growth (Figure 1). To address this gap, we present the first systematic literature review of multiplicity in machine learning, consolidating existing discussions and identifying overarching trends.

To ensure comprehensive coverage, we search across various online repositories (DBLP & ACM Digital Library) using multiple search terms (*'rashomon'*, *'model multiplicity'*, *'set of good models'*), followed by rigorous manual filtering. We eventually found 80 papers that deeply engaged with multiplicity as a central theme in their contributions. Each paper was also manually tagged with all applicable tags, and relevant statistics are presented in Figure 1. The growing interest in the field is evident, with literature on a wide range of problems related to multiplicity. Precise details about the literature review process are in the Appendix (§A).

Contributions. Building on the insights from our literature review, we make the following contributions to multiplicity literature. First, we revisit the Rashomon effect, emphasizing the role of developer choices, and propose a novel *Intent-Convention-Arbitrariness* (ICA) framework to provide formal foundations for future ethnographic studies (§2). Expanding on this, we extend the definitions of Rashomon sets and multiplicity to include underrepresented and unexplored directions of research, anticipating several future subdomains of multiplicity (§3). Next, we formally distin-

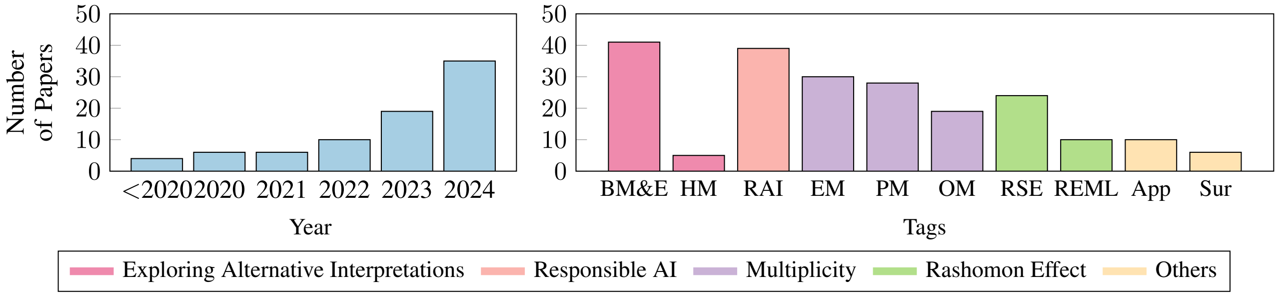

Figure 1: Systematic review of the number of papers over the years and their categorization. Each paper can have multiple tags, marking all categories of contributions made by the paper. Details of the tags are in the Appendix (§A). **BM&E:** Better Models & Ensembles; **HE:** Hacking Metrics; **RAI:** Responsible AI; **PM:** Predictive Multiplicity; **EM:** Explanation Multiplicity; **OM:** Other Multiplicity; **RSE:** Rashomon Set Exploration; **REML:** Rashomon Effect in ML; **App:** Application; **Sur:** Survey.

guish multiplicity from related concepts of uncertainty and variance; and provide both mathematically grounded differences as well as practical guidance on when to adopt each perspective (§4). Finally, we trace two overarching trends in the multiplicity literature: its role in exploring diverse interpretations during model selection (§5), and its broader implications within responsible AI (§6). We conclude by identifying open research questions to encourage future work.

2 The Rashomon Effect in Machine Learning

Taking its name from Akira Kurosawa's 1950 film *Rashomon*, the Rashomon effect is an epistemological framework that highlights the subjectivity and ambiguity inherent in human perception (Anderson 2016; Davis, Anderson, and Walls 2015). Borrowing from Davis, Anderson, and Walls (2015), the Rashomon effect can be defined as *"a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, [...]"*. The Rashomon effect has been studied in several different domains, like the influence of cognitive biases on memory (Tindale 2016; Trabasso 2018), the impact of culture and the fluidity of truth in ethnographic studies (Heider 1988), the study of context, medium, and framing of communication (Anderson 2016; Soreanu and German 2022), the unreliability of eyewitnesses (Pansky, Koriat, and Goldsmith 2005; Hirst and Brown 2011), and–central to our discussion–algorithmic modelling and machine learning (Breiman 2001; Black, Raghavan, and Barocas 2022).

The term Rashomon effect was first introduced into algorithmic modelling by Breiman (2001), pointing out the presence of a set of good models that all achieve similar error rates. It has since been used in discussions of statistical modelling (Bonate 2006; Ueki and Kawasaki 2013), null hacking (Protzko 2018), designing robust algorithms (Tulabandhula and Rudin 2014; Castillo et al. 2008), measuring variable importance (Dong and Rudin 2019; Fisher, Rudin, and Dominici 2019), and applications in various domains (Kang et al. 2018; Chantre et al. 2018). More recently, it has found a resurgence with increasing attention given to *multiplicity* in machine learning, evident both in studies that directly address the topic (Marx, Calmon, and Ustun 2020; Black, Raghavan, and Barocas 2022; Rudin et al. 2024; Del Giudice 2024; Biecek et al. 2024) and in research that situates multiplicity within the broader context of other fields (Møllersen and Holsbø 2023; Rudin et al. 2022; Molnar et al. 2020; Jiang et al. 2024b; Biecek and Samek 2024).

Why do we see the Rashomon effect? In machine learning, data serves as a proxy for the real world, yet it inherently loses information at multiple stages. The first step—translating the world into a data generation process—simplifies complex relationships, introducing randomness to account for uncontrollable aspects. Zhang et al. (2020) termed this "distributional complexity" (associated with 'aleatoric'–Latin *aleatorius*–meaning "dice" or "game of chance"), reflecting the challenge of how well the distribution represents the real world. However, even this distribution remains out of our reach; instead, we work with finite samples. This second step of information loss, described as "approximation complexity" (associated with 'epistemic'–Latin *episteme*–meaning "knowledge") (Zhang et al. 2020), relates to how well we can approximate the underlying generation using finite data. Unlike distributional complexity, which is irreducible, approximation complexity can be mitigated through better data quality and improved algorithms.

Together, distributional and approximation complexities define the fundamental loss of information in learning, resulting in gaps where multiple interpretations, i.e., the Rashomon effect, can arise. While the lens of information loss is insightful, it is limiting in its ability to provide an operational framework to address these challenges. Therefore, we instead focus on the role of developer choices in model design, laying the foundation for discussions on multiplicity.

2.1 Design Choices and Model Selection

Designing a machine learning model involves a series of interconnected choices. Beginning with the data, decisions are made regarding how to process and filter data, which features to select, etc. (Meyer, Albarghouthi, and D'Antoni 2023; Simson, Pfisterer, and Kern 2024; Cavus and Biecek 2024b) Beyond data, the learning algorithm design further entails numerous decisions: model architecture (Arnold et al. 2024; Rudin et al. 2024), hyperparameters (Bouthillier

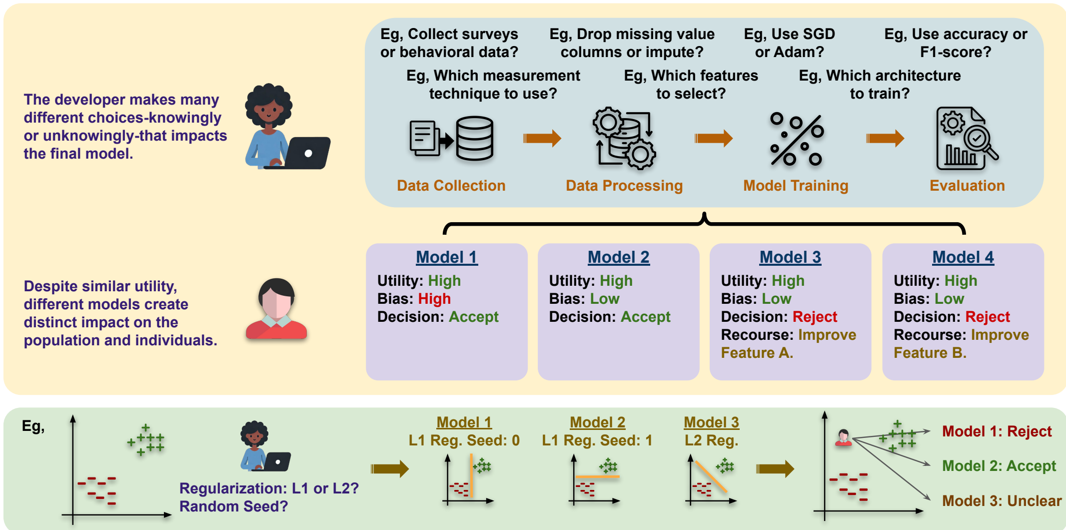

Figure 2: Impact of developer choices on individuals and the population downstream.

et al. 2021; Arnold et al. 2024), various forms of stochasticity (Picard 2021; Bouthillier et al. 2021; Pecher, Srba, and Bielikova 2024; Sellam et al. 2022; McCoy, Min, and Linzen 2020), and even the evaluation and model selection criteria (Ganesh et al. 2024). Each decision contributes to the cascade of choices that directly impacts the multiplicity of the trained models (see Figure 2).

Notably, these choices are not always well-informed. In some cases, they are *intentional*, guided by insights from the literature on the effects of algorithm design on model behaviour (Wu et al. 2021; Ponomareva et al. 2023; Ganesh 2024). In others, they are *conventional*, driven by popular trends or convenience (Shwartz-Ziv and Armon 2022; Dubey, Singh, and Chaudhuri 2022; Creel and Hellman 2022). Finally, some choices remain *arbitrary*, like choosing a random seed. It is through training multiple models and evaluating them that we grasp the impact of these arbitrary choices. To connect these choices with key subdomains in multiplicity, we introduce the *Intent-Convention-Arbitrariness (ICA)* framework. More specifically, we argue that these choices exhibit the following properties:

- **Intentional Choices and Steering Model Behaviour.** We define *intentional* choices as deliberate choices made with an understanding of their impact to achieve desired outcomes. Examples include incorporating bias-mitigating regularization to enhance fairness (Hort et al. 2024; Chen et al. 2023), using simpler models for better interpretability (Rudin et al. 2024), or applying data augmentation to improve robustness (Rebuffi et al. 2021; Eghbal-zadeh et al. 2024). These choices steer model behaviour, giving the developer control over navigating the Rashomon set without the need to train multiple models. Intentional choices are typically informed by extensive prior research or other advancements. Note that choices like selecting a pre-trained model because it is the only available option do not qualify as an intentional choice. Although the developer is aware of the impact of their choice, the decision is made out of necessity rather than deliberate intent.
- **Conventional Choices and Homogenization.** We define *conventional* choices as choices made without knowledge of their impact, out of convenience, or due to lack of alternatives. Examples include adopting popular models or hyperparameters without evaluating their suitability for the specific application (Shwartz-Ziv and Armon 2022; Dubey, Singh, and Chaudhuri 2022), such as using neural networks where simpler models would suffice (Rudin 2019; Shwartz-Ziv and Armon 2022), or applying out-of-the-box fairness, robustness, or explanation techniques without understanding their implications (Gichoya et al. 2023; Rudin 2019; Adebayo et al. 2018; Lipton 2018; Tramer et al. 2020; Carlini et al. 2019). By definition, conventional choices follow established norms or trends within the field rather than addressing specific needs. As a result, these choices contribute to "homogenization", where models trained by different developers exhibit similar behaviour, and can introduce systemic harm (Creel and Hellman 2022; Kleinberg and Raghavan 2021; Bommasani et al. 2022). As shown by Bommasani et al. (2022), even shared components, for instance, some common choices across developers, can lead to homogenized outcomes across multiple systems (§6.2).
- **Arbitrary Choices and Model Selection.** Unlike intentional or conventional choices, *arbitrary* choices have an indeterminate relationship with the final model, usually evaluated after training. Examples include choosing the random seed or arbitrary hyperparameter variations. When dealing with arbitrary choices, making appropriate choices depends on model selection post-training. This has been widely recommended for auditing multiplicity, creating ensembles, or navigating among competing choices (Kulynych et al. 2023; Long et al. 2024; Creel and Hellman 2022). However, training multiple models can be expensive, especially for large models or complex

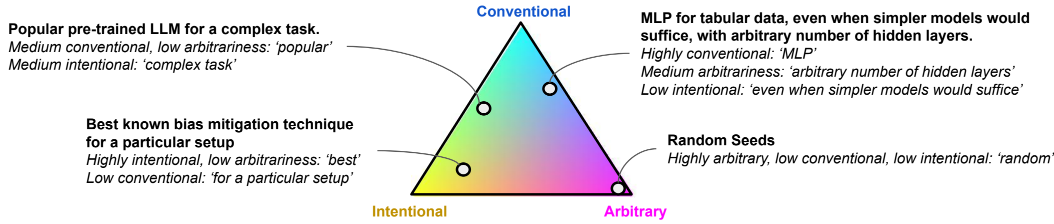

Figure 3: The *intent-convention-arbitrariness* (ICA) framework with some examples.

hypothesis classes (Hsu et al. 2024b,a; Kissel and Mentch 2024; Zhong et al. 2024). Even when feasible, model selection risks overfitting, potentially undermining generalization (Ganesh 2024; Cooper et al. 2021). These challenges don't necessarily dismiss the efficacy of training multiple models. Rather, they highlight the complexities of this approach and motivate a deeper investigation into the benefits and pitfalls of model selection (§5).

In practice, few choices, if any, fall entirely into a single category. Instead, *every choice made by a developer involves a mix of intentional, conventional, and arbitrary factors*. Understanding this balance is essential for navigating the challenges posed by the Rashomon effect (see Figure 3 for some examples). Our ICA framework also lays a foundation for future ethnographic studies exploring design choices in ML. Understanding the model development culture is essential to uncover homogenization or arbitrariness introduced into the system. Our framework provides both the language and conceptual groundwork necessary for such studies.

2.2 Impact of Rashomon Effect

With a grasp of how the Rashomon effect manifests in machine learning, we turn to its impact, i.e., we introduce multiplicity. While the existence of multiple good models is undeniably intriguing, the Rashomon sets should not be a mere academic curiosity. These sets only matter in context, where a change of perspective by choosing a different model from the set, i.e., a different developer choice, influences real-world outcomes (see example in Figure 2). Multiplicity is, thus, the variation in model behaviour across the Rashomon set that holds contextual value. Here, we provide a brief intuition of the various forms of multiplicity, setting the stage for its formal definition in the next section.

The most straightforward example of multiplicity is conflicting predictions from models in the Rashomon set, known as predictive multiplicity (Marx, Calmon, and Ustun 2020). Such conflicts create arbitrariness in decision-making, undermining the reliability of these models and hampering effective planning (D'Amour et al. 2022; Cooper, Frankle, and De Sa 2022; Watson-Daniels et al. 2024; Milani Fard et al. 2016). While predictive multiplicity can be harmful in critical domains, such as medical or legal decisions, it is, however, not inherently bad. For instance, purposefully controlled arbitrariness (called 'randomness' to distinguish from uncontrolled arbitrariness) can help address the concerns of outcome homogenization (Jain et al. 2024; Jain, Creel, and Wilson 2024; Creel and Hellman 2022; Barocas, Hardt, and Narayanan 2023).

Unsurprisingly, predictive multiplicity has received significant attention in the literature (35% of papers in our systematic review; see Figure 1). Yet, this is only one aspect of the Rashomon effect. Consider, for instance, the inconsistency in explanations provided by models within the Rashomon set. Studies have shown that models in the Rashomon set often produce conflicting feature attribution scores (Laberge et al. 2023; Li, Barnard, and Deng 2024; Gunasekaran, Mistry, and Chen 2024; Müller et al. 2023; Poiret et al. 2023; Okazaki et al. 2024; Tan 2024; Kumar et al. 2023), which can undermine trust, e.g., confusing clinicians during AI-assisted diagnostics. Similarly, counterfactuals generated by one model in the Rashomon set often fail to transfer to others (Pawelczyk, Broelemann, and Kasneci 2020; Leofante, Botoeva, and Rajani 2023; Jiang et al. 2024a; Hasan and Talbert 2022). This poses significant challenges for algorithmic recourse when models are regularly updated, as recourse provided by one model may become invalid when replaced by another, questioning their legitimacy and undermining user trust (Rawal, Kamar, and Lakkaraju 2021; Leofante, Botoeva, and Rajani 2023).

More broadly, multiplicity is any form of behavioural difference between models in the context of real-world consequences. These effects are multifaceted, and while predictive and explanation multiplicities are the most recognized, narrowing our focus underestimates the broader risks associated with other underrepresented forms of multiplicity (only $\sim 24\%$ of papers in our systematic review cover other forms of multiplicity; see Figure 1). In the next section, we will formally define both Rashomon sets and multiplicity, expanding upon existing definitions in the literature.

3 Definitions and Metrics

We now define Rashomon sets and multiplicity, building on existing literature while expanding the scope to encompass a wider range of works. Additionally, we review multiplicity metrics and the literature on measuring multiplicity.

3.1 Formalizing Rashomon Sets and Multiplicity

The concept of multiplicity is deeply rooted in the Rashomon effect, and the models illustrating this Rashomon effect are together known as a Rashomon set, a set of competing models, a set of good models, ϵ-Rashomon set, ϵ-Level set, etc. We'll stick with the term Rashomon set for consistency. Rashomon set represents a set of models that are practically indistinguishable, underscoring the arbitrariness in choosing one model over another. Thus, we need to begin by defining these models, i.e., the Rashomon set.

Existing work in multiplicity tends to adopt a narrow view of these models. Much of the research that formalizes the multiplicity problem restricts model choices to a specific hypothesis class $\mathcal{H}$ and/or limits them to training on a fixed, pre-processed dataset $\mathcal{D}$ (Marx, Calmon, and Ustun 2020; Teney, Peyrard, and Abbasnejad 2022; Watson-Daniels et al. 2023). However, this overlooks developer choices made during data collection, data processing, and even model training, all of which can influence the final model, as discussed above. Similarly, most existing studies define indistinguishability solely in terms of loss (Teney, Peyrard, and Abbasnejad 2022; Paes et al. 2023; Du, Ngo, and Wu 2024; Hamman et al. 2024), disregarding other choices involved in designing evaluation criteria and model selection.

To explicitly broaden the definition of Rashomon sets, we introduce a set of metric delta functions, $\mathbf{\Delta}^P$, and corresponding thresholds $\mathcal{E}^P$. A metric delta function takes as input two models and measures the difference between them under the given metric. These metric deltas determine whether two models are indistinguishable: if the difference in performance for every $\delta_i^P \in \mathbf{\Delta}^P$ falls within the corresponding threshold $\epsilon_i^P \in \mathcal{E}^P$. For example, one might define the metric delta for accuracy as $\delta_{Acc,D}(h_1, h_2) = |Acc(h_1, D) - Acc(h_2, D)|$. Formally,

Definition 1 (Rashomon Set) *Two models h_1, h_2 belong to the same Rashomon set under performance constraints $(\mathbf{\Delta}^P, \mathcal{E}^P)$ if they exhibit similar performance for every metric in the given performance constraints, i.e.:*

$$\delta_i^P(h_1, h_2) \leq \epsilon_i^P \quad \forall\, (\delta_i^P, \epsilon_i^P) \in (\mathbf{\Delta}^P, \mathcal{E}^P) \tag{1}$$

Next, we turn to defining the context in which these models exhibit diverse behaviour, i.e., multiplicity. Again, as previously noted, most existing work focuses on either predictive or explanation multiplicity (see Figure 1), with limited attention given to other forms such as fairness multiplicity (Islam, Pan, and Foulds 2021; Coston, Rambachan, and Chouldechova 2021; Ganesh et al. 2023; Gillis, Meursault, and Ustun 2024; Wang and Russakovsky 2021), allocation multiplicity (Jain et al. 2025), OOD robustness multiplicity (Teney, Peyrard, and Abbasnejad 2022; McCoy, Min, and Linzen 2020), model complexity multiplicity (Semenova et al. 2024; Liu et al. 2022; Rudin et al. 2024; Semenova, Rudin, and Parr 2022; Coupkova and Boutin 2024), and feature interaction multiplicity (Li et al. 2024), among others. To capture the various impacts, we generalize multiplicity by binding it to a metric delta δ^M and corresponding threshold ϵ^M, on models belonging to the Rashomon set. Formally,

Definition 2 (Model Multiplicity) *Two models h_1, h_2 exhibit multiplicity under performance constraints $(\mathbf{\Delta}^P, \mathcal{E}^P)$ and multiplicity constraint (δ^M, ϵ^M), if they have similar performance for every metric in the performance constraints yet differ on the metric in the multiplicity constraint, i.e.:*

$$\begin{aligned} &\delta_i^P(h_1, h_2) \leq \epsilon_i^P \quad \forall\, (\delta_i^P, \epsilon_i^P) \in (\mathbf{\Delta}^P, \mathcal{E}^P) \\ \textit{and} \quad &\delta^M(h_1, h_2) > \epsilon^M \end{aligned} \tag{2}$$

Although our definitions are quite similar to existing literature, we have deliberately generalized them to encompass a broader range of underrepresented works. These are not radical changes, but we believe are crucial in drawing attention to various overlooked choices in model design and to better understand their role in multiplicity. However, we also recognize that in certain contexts, a specific definition of multiplicity might be needed. In such contexts, our definition can be reduced appropriately to match the use case.

3.2 Evaluating Multiplicity

We compile a comprehensive list of metrics from the literature that measure various forms of multiplicity in Table 1. For each metric, we mark its original objective, the problem setting, resolution (i.e., whether the metric applies to individual data points or the entire dataset), and if it exhibits set monotonicity within the Rashomon set (i.e., whether a reduction in the Rashomon set size implies a monotonic change in the metric). Monotonicity in a metric can be desirable in certain scenarios (Hsu and Calmon 2022), because it ensures that reducing the Rashomon set size will either decrease or maintain the multiplicity. By systematically recording different facets of each metric in Table 1, we hope to provide a structure for practitioners to identify the most suitable metric for their specific needs.

We also revisit the challenge of training multiple models to evaluate multiplicity, exploring more efficient alternatives. Madras, Atwood, and D'Amour (2019) propose a local ensembling technique to quantify underspecification by analyzing the loss curvature, eliminating the need to train multiple models. Hsu et al. (2024b) use a similar local approach and show that Monte Carlo dropout can be adopted to approximate multiplicity when constrained by utility considerations (i.e., Rashomon sets). In contrast to these local methods, Kissel and Mentch (2024) introduce a model path selection technique that incrementally builds from simpler to more complex models. This approach efficiently constructs the Rashomon set by recursively increasing the complexity of plausible models. Other model class-specific techniques have also been proposed to explore Rashomon sets more efficiently (Hsu et al. 2024a; Mata, Kanamori, and Arimura 2022; Zhong et al. 2024; Xin et al. 2022). Despite these advancements, much work remains to be done to make the enumeration of Rashomon sets more efficient and practical.

4 Multiplicity, Uncertainty and Bias-Variance Decomposition

When examining arbitrariness in decision-making, machine learning research often focuses on prediction uncertainty (Gal 2016; Smith et al. 2024) or the bias-variance decomposition (Domingos 2000; Kong and Dietterich 1995; Brieman 1996). With extensive literature already present in these areas, a natural question arises: *What unique perspectives does multiplicity bring that is not already covered by these concepts?* In this section, we formalize the interplay between multiplicity, uncertainty, and bias-variance decomposition, addressing this question both mathematically and through practical recommendations.

Metric	Original Objective	Problem Setting	Monotonic	Resolution
Ambiguity	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Obscurity	Predictive Multiplicity	Multi-Class Classification	No	Dataset
Discrepancy	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Degree of Underspecification	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Viable Prediction Range	Predictive Multiplicity	Probabilistic Classification	Yes	Individual
Rashomon Capacity	Predictive Multiplicity	Probabilistic Classification	Yes	Individual
Multi-target Ambiguity	Predictive Multiplicity	Multi-Target Classification	Yes	Dataset
Rank List Sensitivity	Predictive Multiplicity	Recommender Systems	-	Dataset
Std. of Scores	Predictive Multiplicity	Agnostic	No	Dataset
Self-consistency	Arbitrariness	Multi-Class Classification	No	Individual
Representational Multiplicity	Procedural Multiplicity	Agnostic	No	Dataset
Region Similarity Score	Procedural Multiplicity	Agnostic	-	Dataset
ϵ-robust to Dataset Multiplicity	Dataset Multiplicity	Regression	-	Dataset
Unfairness Range	Fairness Multiplicity	Agnostic	Yes	Dataset
Rashomon Ratio	Size of Rashomon Set	Agnostic	Yes	Dataset
Underspecification Score	Underspecification	Multi-Class Classification	No	Individual
Accuracy Under Intervention	Metric Multiplicity	Multi-Class Classification	-	Dataset
Metric	**Original Objective**	**Explanation Technique**	**Monotonic**	**Resolution**
Consistency	Explanation Multiplicity	Agnostic	No	Dataset
Model Class Reliance	Explanation Multiplicity	Model Reliance	Yes	Dataset
Attribution Agreement	Explanation Multiplicity	Feature Attribution	-	Individual
Profile Disparity Index	Explanation Multiplicity	Model Profile	-	Dataset
Inv. Cost of Neg. Surprise	Explanation Multiplicity	Counterfactuals	-	Dataset
Variable Importance Clouds	Explanation Multiplicity	Feature Attribution	Yes	Dataset
Coverage & Interval Width	Explanation Uncertainty	Agnostic	No	Individual

Table 1: Multiplicity metrics along with original objective, problem setting, monotonicity wrt to the Rashomon set, and metric resolution. Metrics with no entry under 'Monotonic' compare only two models. Metric citations in the Appendix (§B).

4.1 Multiplicity and Uncertainty

Prediction Uncertainty: We start by defining uncertainty, drawing heavily from Smith et al. (2024). Prediction uncertainty quantifies the degree of confidence–or lack thereof–in a model's predictions. As it reflects the lack of confidence in a model's predictions, uncertainty is often represented as the randomness (or *entropy*) in those predictions. Formally, prediction uncertainty is commonly defined as:

$$U(x, D) = \mathrm{H}_y[\mathrm{Prob}(y|x, D)] \tag{3}$$

$$= \mathrm{H}_y[\mathrm{E}_{\theta\sim\mathrm{Prob}(\theta|D)}[\mathrm{Prob}(y|x,\theta)]] \tag{4}$$

where x is the input for which we're measuring uncertainty, y is the output, D is the training data, and θ is the parametric representation of models. $U(x, D)$ is the prediction uncertainty, while $\mathrm{H}[.], \mathrm{E}[.], \mathrm{Prob}[.]$ represents entropy, expectation, and probability distribution, respectively. The subscript for each statistical measure specifies the random variable or the distribution on which the measure is calculated.

We redirect the interested reader to the uncertainty literature (Hüllermeier and Waegeman 2021; Tran, Snoek, and Lakshminarayanan 2020; Kendall and Gal 2017; Gal 2016), as we do not expand further here. We simply restate these definitions to compare them with multiplicity.

Predictive Multiplicity through the lens of Uncertainty: We temporarily redefine multiplicity, drawing on the same principles used to define uncertainty. In simple terms, we also define multiplicity as the entropy of predictions, but only limited to models within the Rashomon set R. Thus, we can formalize multiplicity $M(x, D)$ as:

$$M(x, D) = \mathrm{H}_y[\mathrm{E}_{\theta\sim\mathrm{Prob}_R(\theta)}[\mathrm{Prob}(y|x,\theta)]] \tag{5}$$

$$\mathrm{Prob}_R(\theta) = \frac{\mathrm{Prob}(\theta|D)}{\sum\limits_{\theta\in R}\mathrm{Prob}(\theta|D)} \text{ if } \theta\in R; \;\; 0 \text{ otherwise} \tag{6}$$

where $\mathrm{Prob}_R(\theta)$ is a modified probability distribution that only includes the models in the Rashomon set R.

Comparing equations 4 and 5, it is clear that the expectation terms are defined over different distributions: over all possible models for uncertainty (eq 4), and over only models within the Rashomon set for multiplicity (eq 5). In other words, while uncertainty measures the potential variance in model predictions for the complete hypothesis class, multiplicity focuses only on models in the Rashomon set, i.e., has a more practical view towards model selection. But why does this distinction matter, and why should we care about both? To answer this, we discuss practical scenarios where viewing a problem through the lens of multiplicity is more appropriate than uncertainty and vice versa.

Uncertainty or Multiplicity? Choosing the Right Lens. Multiplicity examines prediction consistency, while uncertainty assesses confidence. Uncertainty is better suited for modelling the information-theoretic relationships, while multiplicity, on the other hand, actively explores the various interpretations that can emerge during learning. Similarly, when examining how different modelling choices or model selection criteria can influence outcomes, the lens of multiplicity proves invaluable. We outline some characteristics to look for when deciding between the two.

- *Uncertainty provides an information-centric perspective.* As uncertainty definitions are derived from information theory (Box and Tiao 1973; Kendall and Gal 2017; Gal 2016), it is a fundamentally better fit for related analyses. For example, uncertainty plays a crucial role in active learning, by finding instances most likely to provide maximum new 'information' (Yang and Loog 2016; Sharma and Bilgic 2017; Nguyen, Shaker, and Hüllermeier 2022).
- *Uncertainty is sufficient for distributional complexity.* Noise in real-world data may result in a lack of predictive power to make reliable decisions (Wang et al. 2024). Having access to different interpretations through multiplicity adds little value in such cases.
- *Uncertainty quantification is more efficient, but research in multiplicity quantification is growing rapidly.* Uncertainty is streamlined in modern machine learning pipelines through Bayesian networks and model calibration (Niculescu-Mizil and Caruana 2005; Vaicenavicius et al. 2019), offering a cost-effective alternative to multiplicity. Even when training multiple models, there is no definitive way to ensure that every trained model falls in the Rashomon set. Thus, not all trained models contribute to measuring multiplicity, whereas they are still valuable for quantifying uncertainty. That said, advancements in multiplicity research are already challenging this dynamic (§3.2) and may continue to reshape it in the future.
- *Multiplicity is aligned with learning theory and hierarchical optimization.* Every decision in the learning algorithm influences the underlying optimization. Multiplicity can help scrutinize how each choice shapes the final model. Applications include the impact of data processing, random seeds, hyperparameters, etc. (Islam, Pan, and Foulds 2021; Meyer, Albarghouthi, and D'Antoni 2023; Ganesh 2024; Cavus and Biecek 2024b), or, broadly, any form of bi-level or constrained optimization (Black et al. 2024; Semenova et al. 2024; Sun, Wang, and Rudin 2024; Ganesh et al. 2023; Gerchick et al. 2023).
- *Multiplicity is better suited to explore alternative interpretations.* Choosing among different learned interpretations can introduce arbitrariness. Multiplicity, particularly Rashomon sets, enables exploration of these alternative interpretations. Examples include personalization with model mixtures, combining multiple models, homogenization concerns, etc. (Ma et al. 2020; Creel and Hellman 2022; Wu, Black, and Chandrasekaran 2024; Bommasani et al. 2022; Black, Leino, and Fredrikson 2021; Laberge et al. 2023; Breiman 2001; Long et al. 2024)

Note that our recommendations paint a broad picture of when the lens of multiplicity or uncertainty could be useful, but these are intended only as guidelines, and deviation from these may be warranted in specific contexts.

4.2 Multiplicity and Variance

Another measure of arbitrariness in decisions is the 'variance' from the bias-variance decomposition. Error in machine learning is categorized into three parts: irreducible error, bias, and variance (Domingos 2000; Brieman 1996). The terms 'bias' and 'variance' describe how well the model approximates the underlying distribution, i.e., approximation complexity, where 'bias' captures how well the chosen model fits the given data, while 'variance' reflects the model's sensitivity to variations in the dataset.

At first glance, 'variance' might seem like a natural way to measure arbitrariness in decision-making. However, bias-variance decomposition is typically confined to a single model, focusing only on sensitivity to the underlying dataset. While this is valuable, guiding preferences for models with lower bias or variance, it does not address the broader arbitrariness introduced by the entire learning pipeline. Multiplicity, in contrast, enables comparisons across different models generated through the pipeline. Formally,

$$Error = IrreducibleError + Bias + \mathbf{Var}_{\mathbf{D}}[\mathbf{f}_\theta(\mathbf{x})] \quad (7)$$

$$PredictiveMultiplicity = \mathbf{Var}_{\theta \sim \mathrm{Prob}_{\mathbf{R}}(\theta|\mathbf{D})}[\mathbf{f}_\theta(\mathbf{x})] \quad (8)$$

where $f_\theta(.)$ is the predictive function for the learned parameter value θ, and $Var[.]$ represents variance. We use a metric of predictive multiplicity that also uses variance to quantify multiplicity, facilitating clearer comparisons (Heljakka et al. 2022; Long et al. 2024; Hamman et al. 2024). While the concept of 'variance' from bias-variance decomposition focuses on a single model's sensitivity to variations in the dataset, multiplicity, in contrast, captures changing predictions across multiple models. Black, Raghavan, and Barocas (2022) provides a similar comparison, bounding the expected disagreement between two models (potentially quantifying multiplicity) using the expected variance.

We showed that multiplicity is neither universally redundant nor superior when compared to uncertainty or bias-variance decomposition. Instead, we offer guidance on when each approach is most appropriate. The interplay between multiplicity, uncertainty, and bias-variance decomposition remains complex, underscoring the need for further research to better understand and effectively utilize these concepts.

5 Exploring Alternative Interpretations

Now that we've discussed the formalization of multiplicity, let's delve into its real-world implications. One of the biggest advantages of multiplicity lies in its ability to explore various 'good' learned interpretations (Black, Raghavan, and Barocas 2022; Rudin et al. 2024). When multiple interpretations exist, it is reasonable to expect that some of them may exhibit certain desirable properties, such as better fairness, robustness, interpretability, etc. In this section, we study how multiplicity facilitates such exploration and its broader implications. Our discussion builds on what Black, Raghavan, and Barocas (2022) refer to as the *aggregate*-level effects of multiplicity, while incorporating insights from more recent developments in the field.

5.1 Searching Instead of Optimizing

Machine learning often involves tackling complex optimizations, including two common hierarchical optimization problems: *bi-level optimization* and *constrained optimization*. Bi-level optimization refers to scenarios where one optimization problem depends on variables governed by another nested optimization (Sinha, Malo, and Deb 2017; Zhang et al. 2024). A classic example is hyperparameter optimization. Constrained optimization, on the other hand, involves solving the optimization problems under specific constraints (Bertsekas 2014; Goodfellow, Shlens, and Szegedy 2014). Examples include enforcing constraints of fairness, robustness, safety, etc. (Gallego-Posada 2024; Dwork et al. 2012). These optimization problems can be notoriously difficult to solve. Challenges arise from the complexity of formalizing constraints, the difficulty of creating differentiable relaxations, the absence of closed-form solutions, or evolving client requirements, among others.

Interestingly, multiplicity offers a practical workaround: *if you can't optimize, search for it!* Brute-force strategies of searching through various potential solutions are well-established in both bi-level optimization and constrained optimization literature (Sinha, Malo, and Deb 2017; Zhang et al. 2024; Gallego-Posada 2024). Though more efficient reductions and mathematical alternatives are preferred when feasible, they are not always possible. In such cases, searching through potential solutions to find the best fit becomes a viable strategy, and multiplicity plays a pivotal role in this search (see Figure 4). Multiplicity has been used in the literature to find models with lower bias (Black et al. 2024; Simson, Pfisterer, and Kern 2024; Ganesh et al. 2024; Islam, Pan, and Foulds 2021), smaller model complexity (Semenova, Rudin, and Parr 2022; Semenova et al. 2024; Boner et al. 2024; Rudin et al. 2024; Doctor, Mao, and Mhaskar 2024; Coupkova and Boutin 2024), better explanations (Sun, Wang, and Rudin 2024; Veran, Portier, and Fouquet 2023), improved generalizability (Li, Barnard, and Deng 2024), and the ability to allow personalization (Ma et al. 2020). Beyond simply enumerating the Rashomon set to search for better models, several recent works have also shown how the visualization of the Rashomon set can empower developers to select models that meet their specific requirements (He and Shaposhnik 2023; Eerlings et al. 2024; Luyten et al. 2024).

5.2 Ensembles and More

Selecting the "best" model may not always be recommended, particularly when no single interpretation of the data can be optimal. Instead, combining insights from multiple models is preferred. Techniques like prediction ensembling, or bagging, have long been a central recommendation for stability in machine learning (Breiman 1996; Dietterich 2000; Long et al. 2024). Many methods in the multiplicity literature have capitalized on combining models in the Rashomon set. While literature in this direction primarily focuses on aggregating model explanations to create more stable and reliable explanations (Jiang et al. 2024a; Sreedharan, Kambhampati et al. 2018; Baniecki, Parzych, and Biecek 2024; Donnelly et al. 2023; Fisher, Rudin, and Dominici 2019; Dong and Rudin 2020; Kobylińska et al. 2024; Zuin et al. 2021; Hamamoto and Egi 2021; Kowal 2022), other works have also shown the benefits of aggregating fairness scores (Coston, Rambachan, and Chouldechova 2021), individual probabilities (Roth, Tolbert, and Weinstein 2023), or regression analysis to discover causality (Ueki and Kawasaki 2013). These techniques demonstrate the value of leveraging multiplicity not just to select a single best model but instead to combine multiple learned interpretations.

5.3 Hacking Metrics with Multiplicity

While multiplicity can discover better models, it also introduces risks, particularly the potential of exploiting these methods to circumvent regulatory requirements and interventions. This is more prevalent when broad principles, such as fairness, are reduced to specific benchmarks or metrics (Black, Gillis, and Hall 2024). By leveraging multiplicity, it becomes possible to 'hack' these metrics, producing models that meet the specified criteria without truly adhering to the underlying principles (Coker, Rudin, and King 2021; Cooper et al. 2021; Møllersen and Holsbø 2023). Several studies in the literature have shown that such a search can produce models capable of "regulatory-washing", being able to manipulate explanations (Sharma et al. 2024; Aïvodji et al. 2021; Shahin Shamsabadi et al. 2022) and fairness scores (Forde et al. 2021; Black, Gillis, and Hall 2024; Ganesh et al. 2024). Such manipulation can also occur unintentionally as a result of overfitting to a given metric (Black, Gillis, and Hall 2024; Cooper et al. 2021; Ganesh 2024), underscoring the need for vigilance against the misuse of multiplicity and calling for a more robust operationalization of regulatory frameworks (Black, Gillis, and Hall 2024).

6 Multiplicity and Responsible AI

We now place multiplicity in the broader landscape of responsible AI. In this section, we examine the two key concerns for individuals originating from multiplicity, i.e., arbitrariness in model selection and outcome homogenization. Again, our discussion here builds on what was termed as the *individual*-level effects of multiplicity by Black, Raghavan, and Barocas (2022), while also focusing on topics that have gained more interest recently, such as homogenization.

6.1 Arbitrariness as a Responsible AI Concern

Arbitrary decisions in an automated system can be deeply concerning when they can have direct and lasting impacts on human lives (Black, Raghavan, and Barocas 2022; Gomez et al. 2024; Sokol et al. 2024; Watson-Daniels et al. 2024). Borrowing an analogy from Gomez et al. (2024), imagine a judge deciding legal cases by flipping a coin. While this may seem extreme, it demonstrates how models can have arbitrariness embedded in them due to an analogous coin flip done by the developer during model design. This aligns closely with our ICA framework (§2.1), where we discuss the *arbitrary* choices that can contribute to multiplicity.

There are contexts where a degree of "controlled randomness" may be acceptable—or even necessary (§6.2). However, arbitrariness is a significant concern in scenarios

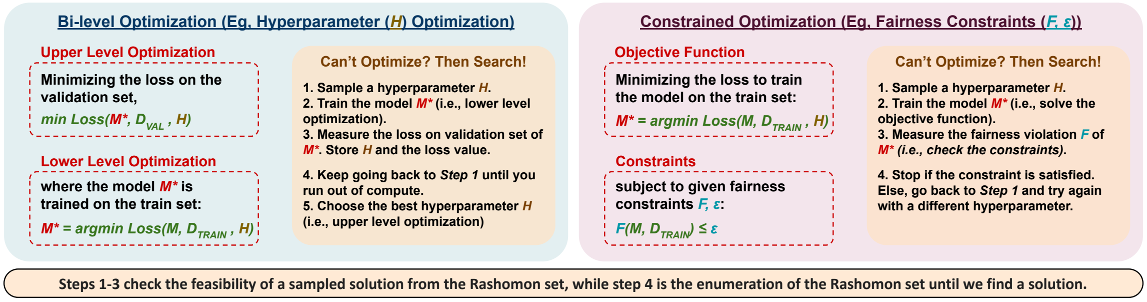

Figure 4: The role of multiplicity in brute-force search for bi-level and constrained optimization problems in machine learning.

where individuals lack access to other *'equivalent opportunities'* (Barocas, Hardt, and Narayanan 2023). For instance, in hiring, some level of arbitrariness may be acceptable, or even necessary to deal with the concerns of homogenization (§6.2). This is because individuals looking for a job often seek multiple opportunities across companies, increasing their chances of being hired elsewhere. In contrast, domains such as law or medicine typically involve singular, high-stakes decisions with no equivalent alternatives. In such situations, the presence of arbitrariness raises serious concerns for the responsible deployment of machine learning models.

The negative effects of multiplicity extend beyond the arbitrariness of just the final prediction (§2.2). For instance, multiplicity in counterfactual explanations can impact the validity of algorithmic recourse (Hasan and Talbert 2022; Leventi-Peetz and Weber 2022; Baniecki, Parzych, and Biecek 2024; Hamman et al. 2023; Leofante, Botoeva, and Rajani 2023; Jiang et al. 2024a). The feasibility or nature of recourse might hinge on *arbitrary* design choices made during model development. These decisions can have real-world implications; for instance, a recourse provided by one model may become invalid if the model is updated, invalidating previous efforts. This inconsistency raises ethical and legal concerns (Kobylińska et al. 2024; De Toni et al. 2024).

This arbitrariness becomes even more problematic when it disproportionately impacts different individuals, particularly harming underrepresented demographics. As hinted earlier, a significant source of arbitrariness is the model's lack of ability to learn the underlying distribution. Underrepresented groups often face these information gaps, which can manifest as data scarcity due to limited historical records or a lack of understanding of cultural context for how data relates to predictions (Barocas, Hardt, and Narayanan 2023; Hooker 2021; Buolamwini and Gebru 2018; Birhane 2022; Geirhos et al. 2020). Research has consistently shown that such disparities exacerbate existing inequalities—whether through arbitrariness, uncertainty, or multiplicity (Farnadi, Havaei, and Rostamzadeh 2024). The resulting disproportionate harms across groups highlight the pressing need to address arbitrariness in critical domains (Gomez et al. 2024; Cooper et al. 2024; Ali, Lahoti, and Gummadi 2021).

Responsible AI Constraints and Multiplicity: We saw that multiplicity is a critical consideration in responsible AI. Interestingly, several works have also explored how multiplicity interacts with other pillars of responsible AI. Studies have shown that imposing fairness constraints can inadvertently increase multiplicity (Long et al. 2024; Cavus and Biecek 2024a). However, Long et al. (2024) argue that multiplicity stands outside the fairness-utility trade-off, meaning improvements in fairness do not have to entail increased multiplicity. They demonstrate that multiplicity can often be reduced through techniques like ensembling, while maintaining fairness. Similarly, Kulynych et al. (2023) explore the interaction between differential privacy and multiplicity, revealing that introducing privacy constraints tends to increase multiplicity. These findings underscore the complex interplay between multiplicity and other responsible AI principles, highlighting the need for further research to understand and navigate these trade-offs effectively.

6.2 Multiplicity and Homogenization

As discussed above, the extent to which arbitrariness is problematic often depends on the context, and in certain cases, it might not be a standalone concern. For instance, Creel and Hellman (2022) argue that arbitrariness in hiring decisions, by itself, is neither a legal nor moral issue. Instead, they suggest that the absence of arbitrariness across systems could lead to a different concern, creating an 'algorithmic leviathan', i.e., the standardization of a single outcome across an entire sector. Kleinberg and Raghavan (2021) discuss a similar concern in the form of 'algorithmic monoculture', which would be particularly problematic in interconnected systems, for instance, when multiple banks assess an individual's creditworthiness, algorithmic monoculture would imply that an individual rejected from one bank would be rejected from all banks. Please refer to Table 2 for an overview of other common terms used in this literature.

This phenomenon, known as outcome homogenization, refers to the convergence of decisions due to common design choices across multiple models. In our ICA framework (§2.1), we had termed these as *conventional* choices. Bommasani et al. (2022) shows that outcome homogenization can occur even when different algorithms share only certain components, i.e., homogenization can occur even when only some design choices are *conventional*.

Introducing controlled randomness can mitigate these risks by preventing monocultures. In contexts where arbitrariness in individual decisions is less concerning than homogenization, controlled multiplicity is desirable (Barocas, Hardt, and Narayanan 2023; Perry and Zarsky 2015; Jain,

Term	Definition
Algorithmic Blackballing (Ajunwa 2020)	*'A worker's lack of control over the portability of applicant data captured by automated hiring systems [. . .] raising the specter of an algorithmically permanently excluded class of job applicants'*
Algorithmic Monoculture (Kleinberg and Raghavan 2021)	*'The notion that choices and preferences will become homogeneous in the face of algorithmic curation'*
Algorithmic Leviathan (Creel and Hellman 2022)	*'Automated decision-making systems that make uniform judgments across broad swathes of a sector.'*
Outcome Homogenization (Bommasani et al. 2022)	*'The phenomenon of individuals (or groups) exclusively receiving negative outcomes from all ML models they interact with'*
Generative Monoculture (Wu, Black, and Chandrasekaran 2024)	*'A distribution shift [towards less varied outputs] from source data (i.e., human-generated training data) to model-generated data (i.e., model outputs) for a specific task.'*
Algorithmic Pluralism (Jain et al. 2024)	*'A state of affairs in which the algorithms used for decision-making are not so pervasive and/or strict as to constitute a severe bottleneck on opportunity.'*

Table 2: Various terms used in the homogenization literature.

Creel, and Wilson 2024). In such situations, when an *intentional* design choice is not possible, the developers should prefer *arbitrary* choices over *conventional* ones wherever feasible (§2.1). Despite this idea being widely recognized in the academic literature (Barocas, Hardt, and Narayanan 2023; Perry and Zarsky 2015; Jain, Creel, and Wilson 2024), public perception of intentional randomness in decision-making remains skeptical. A recent study by Meyer et al. (2024) indicates a strong aversion of the end users towards any form of randomization or intentional arbitrariness in automated decision-making. Therefore, fostering greater public awareness about the nuanced impacts of multiplicity is crucial before we can develop and employ potential solutions (Barocas, Hardt, and Narayanan 2023).

7 Open Questions and Emerging Trends

Our systematic survey gives us a unique vantage point to identify and discuss several emerging trends in the field.

- *Expanding the scope of multiplicity beyond predictions and explanations.* Our key motivation for broadening the definition of multiplicity (§3.1) was to incorporate multiplicity beyond predictive and explanatory contexts. Although interest in these aspects is growing, further work is needed to bring them to the forefront. We hope our work will encourage future research that fosters a more holistic understanding of multiplicity's broader impact.
- *Cost-effective enumeration of Rashomon sets.* A major challenge in auditing multiplicity lies in the resource-intensive nature of training multiple models to enumerate the Rashomon set. While we discussed several works that improve the efficiency of this enumeration (§3.2), the need for further research in this direction remains pressing.
- *Mathematical foundations of multiplicity.* Establishing stronger mathematical foundations of multiplicity, for instance, our focus on formalizing the distinction between multiplicity, uncertainty, and bias-variance decomposition (§4), is essential. Fundamental work on the Rashomon effect is less represented (only 12.5%; see Figure 1), highlighting the opportunities for future work on frameworks that rigorously define and explore multiplicity.
- *Multiplicity and its interaction with responsible AI.* Given the conversation of arbitrariness as a concern of responsible model development (§6.1), its interaction with other pillars of responsible AI is warranted. Future research on frameworks that address multiplicity within the broader landscape of responsible AI deployment is needed.
- *Interdisciplinary perspective on multiplicity.* We found many studies in this field do not engage with the concerns of multiplicity in real-world settings. Our systematic review shows that only 47.5% of works explicitly engage with responsible AI concerns (see Figure 1), which we believe to be low given the field's relevance to the responsible deployment of machine learning models. Future collaborative efforts across disciplines are thus crucial.
- *Multiplicity and LLMs.* As models continue to scale, new challenges emerge. For instance, we see increasing attention given to the concerns of monoculture and homogenization (§6.2). Even the evaluation of multiplicity becomes increasingly complex, as training multiple models is often infeasible at this scale. Additionally, we see new dimensions of multiplicity like prompt multiplicity (Ganesh, Shokri, and Farnadi 2025), preference multiplicity, etc., that require deeper examination.

8 Conclusion

In this work, we systemized existing knowledge on multiplicity, uncovering interesting trends. One limitation of our study is the evolving terminology within the field–terms such as "multiplicity" and "Rashomon sets" have gained prominence only recently. As a result, our survey may have missed relevant works that did not explicitly use this terminology. Despite this, our efforts to formalize key discussions–the language around developer choices, definitions, and the distinction between multiplicity, uncertainty, and variance–represent a crucial step toward unifying the field. We also explored broad trends related to the real-world impacts of multiplicity, building from existing literature and highlighting overarching themes that extend beyond their originally studied contexts. We hope our work provides a platform that is both accessible to newcomers and valuable to experts, fostering further research in multiplicity.

Acknowledgements

We would like to thank Shomik Jain for his feedback on an earlier version of the paper. Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, CIFAR Catalyst Grant award, FRQNT and NSERC Discovery Grants program. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Aïvodji, U.; Arai, H.; Gambs, S.; and Hara, S. 2021. Characterizing the risk of fairwashing. *Advances in Neural Information Processing Systems*, 34: 14822–14834.

Ajunwa, I. 2020. An auditing imperative for automated hiring systems. *Harv. JL & Tech.*, 34: 621.

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.

Anderson, R. 2016. The Rashomon effect and communication. *Canadian Journal of Communication*, 41(2): 249–270.

Arnold, C.; Biedebach, L.; Küpfer, A.; and Neunhoeffer, M. 2024. The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*, 12(4): 841–848.

Baniecki, H.; Parzych, D.; and Biecek, P. 2024. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery*, 38(5): 2596–2632.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bayes, T. 1763. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, 53: 370–418. Published posthumously, communicated by Richard Price.

Bertsekas, D. P. 2014. *Constrained optimization and Lagrange multiplier methods*. Academic press.

Biecek, P.; Baniecki, H.; Krzyziński, M.; and Cook, D. 2024. Performance is not enough: the story told by a Rashomon quartet. *Journal of Computational and Graphical Statistics*, (just-accepted): 1–6.

Biecek, P.; and Samek, W. 2024. Position: Explain to Question not to Justify. In *Forty-first International Conference on Machine Learning*.

Birhane, A. 2022. *Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence*. Ph.D. thesis, University College Dublin. School of Computer Science.

Black, E. 2022. *(Un) Fairness Along the AI Pipeline Problems and Solutions*. Ph.D. thesis, Carnegie Mellon University.

Black, E.; Gillis, T.; and Hall, Z. Y. 2024. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 602–615.

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

Black, E.; Leino, K.; and Fredrikson, M. 2021. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–863.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35: 3663–3678.

Bonate, P. L. 2006. *The art of modeling*. Springer.

Boner, Z.; Chen, H.; Semenova, L.; Parr, R.; and Rudin, C. 2024. Using Noise to Infer Aspects of Simplicity Without Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3: 747–769.

Box, G. E. P.; and Tiao, G. C. 1973. Bayesian inference in statistical analysis. *International Statistical Review*, 43: 242.

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24: 123–140.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3): 199–231.

Brieman, L. 1996. Bias, variance, and arcing classifiers. *Univ. of California, Dept. of Stat., TR 460*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

Castillo, E.; Mínguez, R.; Castillo, C.; and Cofino, A. S. 2008. Dealing with the multiplicity of solutions of the l1 and l-inf regression models. *European journal of operational research*, 188(2): 460–484.

Cavus, M.; and Biecek, P. 2024a. An Experimental Study on the Rashomon Effect of Balancing Methods in Imbalanced Classification. *arXiv preprint arXiv:2405.01557*.

Cavus, M.; and Biecek, P. 2024b. Investigating the Impact of Balancing, Filtering, and Complexity on Predictive Multiplicity: A Data-Centric Perspective. *arXiv preprint arXiv:2412.09712*.

Chantre, G. R.; Vigna, M. R.; Renzi, J. P.; and Blanco, A. M. 2018. A flexible and practical approach for real-time weed emergence prediction based on Artificial Neural Networks. *Biosystems engineering*, 170: 51–60.

Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology*, 32(4): 1–30.

Coker, B.; Rudin, C.; and King, G. 2021. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10): 6174–6197.

Cooper, A. F. 2024. *Between Randomness and Arbitrariness: Some Lessons for Reliable Machine Learning at Scale*. Ph.D. thesis, Cornell University.

Cooper, A. F.; Frankle, J.; and De Sa, C. 2022. Non-Determinism and the Lawlessness of Machine Learning Code. In *Proceedings of the 2022 Symposium on Computer Science and Law*, 1–8.

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.

Cooper, A. F.; Lu, Y.; Forde, J.; and De Sa, C. M. 2021. Hyperparameter optimization is deceiving us, and how to stop it. *Advances in Neural Information Processing Systems*, 34: 3081–3095.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

Coupkova, E.; and Boutin, M. 2024. On the Rashomon ratio of infinite hypothesis sets. *arXiv preprint arXiv:2404.17746*.

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. *Journal of Machine Learning Research*, 23(226): 1–61.

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

De Toni, G.; Teso, S.; Lepri, B.; and Passerini, A. 2024. Time Can Invalidate Algorithmic Recourse. *arXiv preprint arXiv:2410.08007*.

Del Giudice, M. 2024. The prediction-explanation fallacy: a pervasive problem in scientific applications of machine learning. *Methodology*, 20(1): 22–46.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Doctor, K.; Mao, T.; and Mhaskar, H. 2024. Encoding of data sets and algorithms. *Applied Numerical Mathematics*, 200: 209–235.

Domingos, P. 2000. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, 231–238. Morgan Kaufmann Stanford.

Dong, J.; and Rudin, C. 2019. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*.

Dong, J.; and Rudin, C. 2020. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12): 810–824.

Donnelly, J.; Katta, S.; Rudin, C.; and Browne, E. 2023. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Advances in Neural Information Processing Systems*, 36: 6267–6279.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580.

Du, A. Y.; Ngo, D. D.; and Wu, Z. S. 2024. Reconciling Model Multiplicity for Downstream Decision Making. *arXiv preprint arXiv:2405.19667*.

Dubey, S. R.; Singh, S. K.; and Chaudhuri, B. B. 2022. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *Neurocomputing*, 503: 92–108.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Eerlings, G.; Vanbrabant, S.; Liesenborgs, J.; Ruiz, G. R.; Vanacken, D.; and Luyten, K. 2024. AI-Spectra: A Visual Dashboard for Model Multiplicity to Enhance Informed and Transparent Decision-Making. *arXiv preprint arXiv:2411.10490*.

Eghbal-zadeh, H.; Zellinger, W.; Pintor, M.; Grosse, K.; Koutini, K.; Moser, B. A.; Biggio, B.; and Widmer, G. 2024. Rethinking data augmentation for adversarial robustness. *Information Sciences*, 654: 119838.

Farnadi, G.; Havaei, M.; and Rostamzadeh, N. 2024. Position: Cracking the Code of Cascading Disparity Towards Marginalized Communities. In *Forty-first International Conference on Machine Learning*.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.

Forde, J. Z.; Cooper, A. F.; Kwegyir-Aggrey, K.; De Sa, C.; and Littman, M. 2021. Model selection's disparate impact in real-world deep learning applications. *arXiv preprint arXiv:2104.00606*.

Gal, Y. 2016. *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.

Gallego-Posada, J. 2024. *Constrained optimization for machine learning: algorithms and applications*. Ph.D. thesis, Université de Montréal.

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*, 4488–4497.

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

Ganesh, P.; Gohar, U.; Cheng, L.; and Farnadi, G. 2024. Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML. *arXiv preprint arXiv:2411.11101*.

Ganesh, P.; Shokri, R.; and Farnadi, G. 2025. Rethinking Hallucinations: Correctness, Consistency, and Prompt Multiplicity. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gerchick, M.; Jegede, T.; Shah, T.; Gutierrez, A.; Beiers, S.; Shemtov, N.; Xu, K.; Samant, A.; and Horowitz, A. 2023. The devil is in the details: Interrogating values embedded in the allegheny family screening tool. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1292–1310.

Gichoya, J. W.; Thomas, K.; Celi, L. A.; Safdar, N.; Banerjee, I.; Banja, J. D.; Seyyed-Kalantari, L.; Trivedi, H.; and Purkayastha, S. 2023. AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology*, 96(1150): 20230023.

Gillis, T. B.; Meursault, V.; and Ustun, B. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 377–387.

Gomez, J. F.; Machado, C.; Paes, L. M.; and Calmon, F. 2024. Algorithmic Arbitrariness in Content Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2234–2253.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gunasekaran, A.; Mistry, P.; and Chen, M. 2024. Which Explanation Should be Selected: A Method Agnostic Model Class Reliance Explanation for Model and Explanation Multiplicity. *SN Computer Science*, 5(5): 1–20.

Hamamoto, M.; and Egi, M. 2021. Model-agnostic ensemble-based explanation correction leveraging rashomon effect. In *2021 IEEE symposium series on computational intelligence (SSCI)*, 01–08. IEEE.

Hamman, F.; Dissanayake, P.; Mishra, S.; Lecue, F.; and Dutta, S. 2024. Quantifying prediction consistency under model multiplicity in tabular LLMs. *arXiv preprint arXiv:2407.04173*.

Hamman, F.; Noorani, E.; Mishra, S.; Magazzeni, D.; and Dutta, S. 2023. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning*, 12351–12367. PMLR.

Hasan, M. G. M. M. 2022. *Targeted Generation of Counterfactual Examples for Data Augmentation and to Mitigate the Rashomon Effect*. Ph.D. thesis, Tennessee Technological University.

Hasan, M. G. M. M.; and Talbert, D. 2022. Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach. In *The International FLAIRS Conference Proceedings*, volume 35.

He, Z.; and Shaposhnik, Y. 2023. Visualizing the Implicit Model Selection Tradeoff. *Journal of Artificial Intelligence Research*, 76: 829–881.

Heider, K. G. 1988. The Rashomon effect: When ethnographers disagree. *American Anthropologist*, 90(1): 73–81.

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

Hirst, W.; and Brown, A. 2011. On the Virtues of an Unreliable Memory. *Grounding sociality: Neurons, mind, and culture*, 95–113.

Hooker, S. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4).

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2): 1–52.

Hsu, H. 2023. *Information-Theoretic Tools for Machine Learning Beyond Accuracy*. Ph.D. thesis, Harvard University Graduate School of Arts and Sciences.

Hsu, H.; Brugere, I.; Sharma, S.; Lecue, F.; and Chen, C.-F. 2024a. RashomonGB: Analyzing the Rashomon Effect and Mitigating Predictive Multiplicity in Gradient Boosting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.; and Chen, C.-F. 2024b. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. In *The Twelfth International Conference on Learning Representations*.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506.

Islam, R.; Pan, S.; and Foulds, J. R. 2021. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 586–596.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.

Jain, S.; Creel, K.; and Wilson, A. 2024. Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. *arXiv preprint arXiv:2404.08592*.

Jain, S.; Suriyakumar, V.; Creel, K.; and Wilson, A. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 197–206.

Jain, S.; Wang, M.; Creel, K.; and Wilson, A. 2025. Allocation Multiplicity: Evaluating the Promises of the Rashomon Set. *arXiv preprint arXiv:2503.16621*.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024a. Recourse under model multiplicity via argumentative ensembling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 954–963.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024b. Robust Counterfactual Explanations in Machine Learning: A Survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 8086–8094. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Kang, J.; Rancati, T.; Lee, S.; Oh, J. H.; Kerns, S. L.; Scott, J. G.; Schwartz, R.; Kim, S.; and Rosenstein, B. S. 2018. Machine learning and radiogenomics: lessons learned and future directions. *Frontiers in oncology*, 8: 228.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kissel, N.; and Mentch, L. 2024. Forward stability and model path selection. *Statistics and Computing*, 34(2): 82.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Kobylińska, K.; Krzyziński, M.; Machowicz, R.; Adamek, M.; and Biecek, P. 2024. Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data. *IEEE Journal of Biomedical and Health Informatics*, 28(11): 6454–6465.

Kong, E. B.; and Dietterich, T. G. 1995. Error-correcting output coding corrects bias and variance. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, 313–321. Morgan Kaufmann Publishers Inc. ISBN 1558603778.

Kowal, D. R. 2022. Bayesian subset selection and variable importance for interpretable prediction and classification. *Journal of Machine Learning Research*, 23(108): 1–38.

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.

Kumar, R.; Koshiyama, A.; da Costa, K.; Kingsman, N.; Tewarrie, M.; Kazim, E.; Roy, A.; Treleaven, P.; and Lovell, Z. 2023. Deep learning model fragility and implications for financial stability and regulation. Technical report, Bank of England.

Laberge, G.; Pequignot, Y.; Mathieu, A.; Khomh, F.; and Marchand, M. 2023. Partial order in chaos: consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364): 1–50.

Leofante, F.; Botoeva, E.; and Rajani, V. 2023. Counterfactual explanations and model multiplicity: a relational verification view. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 19, 763–768.

Leventi-Peetz, A.-M.; and Weber, K. 2022. Rashomon effect and consistency in explainable artificial intelligence (XAI). In *Proceedings of the Future Technologies Conference*, 796–808. Springer.

Li, S.; Barnard, A. S.; and Deng, Q. 2024. Practical attribution guidance for rashomon sets. *arXiv preprint arXiv:2407.18482*.

Li, S.; Wang, R.; Deng, Q.; and Barnard, A. S. 2024. Exploring the cloud of feature interaction scores in a Rashomon set. In *The Twelfth International Conference on Learning Representations*.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Liu, J.; Zhong, C.; Li, B.; Seltzer, M.; and Rudin, C. 2022. FasterRisk: Fast and accurate interpretable risk scores. *Advances in Neural Information Processing Systems*, 35: 17760–17773.

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

Luyten, K.; Eerlings, G.; Liesenborgs, J.; Rovelo Ruiz, G.; Vanbrabant, S.; and Vanacken, D. 2024. Opportunities and Challenges of Model Multiplicity in Interactive Software Systems.

Ma, S.; Schreiner, P. J.; Seaquist, E. R.; Ugurbil, M.; Zmora, R.; and Chow, L. S. 2020. Multiple predictively equivalent risk models for handling missing data at time of prediction: With an application in severe hypoglycemia risk prediction for type 2 diabetes. *Journal of biomedical informatics*, 103: 103379.

Madras, D.; Atwood, J.; and D'Amour, A. 2019. Detecting extrapolation with local ensembles. In *International Conference on Learning Representations*.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Mata, K.; Kanamori, K.; and Arimura, H. 2022. Computing the Collection of Good Models for Rule Lists. In *Proc. the 18th International Conference on Machine Learning and Data Mining (MLDM 2022)*.

McCoy, R. T.; Min, J.; and Linzen, T. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 217–227.

Meyer, A. P.; Albarghouthi, A.; and D'Antoni, L. 2023. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 193–204.

Meyer, A. P.; Kim, Y.-S.; Albarghouthi, A.; and D'Antoni, L. 2024. Perceptions of the Fairness Impacts of Multiplicity in Machine Learning. *arXiv preprint arXiv:2409.12332*.

Milani Fard, M.; Cormier, Q.; Canini, K.; and Gupta, M. 2016. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems*, 29.

Møllersen, K.; and Holsbø, E. 2023. Accounting for multiplicity in machine learning benchmark performance. *arXiv preprint arXiv:2303.07272*.

Molnar, C.; König, G.; Herbinger, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C. A.; Casalicchio, G.; Grosse-Wentrup, M.; and Bischl, B. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 39–68. Springer.

Müller, S.; Toborek, V.; Beckh, K.; Jakobs, M.; Bauckhage, C.; and Welke, P. 2023. An empirical evaluation of the Rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 462–478. Springer.

Nguyen, V.-L.; Shaker, M. H.; and Hüllermeier, E. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1): 89–122.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Okazaki, K.; Ribeiro, T.; Lo, K.-Y.; Sakuma, J.; and Inoue, K. 2024. Data-driven Analysis of Domain Specificity for Explainable Session-based Recommendation System. In *Proceedings of the National Conference of the Japanese Society for Artificial Intelligence, No. 38 (2024)*, 2Q5IS104–2Q5IS104. Japan Society for Artificial Intelligence.

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.

Pansky, A.; Koriat, A.; and Goldsmith, M. 2005. Eyewitness recall and testimony. *Psychology and law: An empirical perspective*, 93–150.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, 809–818. PMLR.

Pecher, B.; Srba, I.; and Bielikova, M. 2024. A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness. *ACM Computing Surveys*, 57(1): 1–40.

Perry, R.; and Zarsky, T. 2015. 'May the Odds Be Ever in Your Favor': Lotteries in Law. *Alabama Law Review*, 66: 1035–1098.

Picard, D. 2021. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*.

Poiret, C.; Grigis, A.; Thomas, J.; and Noulhiane, M. 2023. Can we Agree? On the Rash\= omon Effect and the Reliability of Post-Hoc Explainable AI. *arXiv preprint arXiv:2308.07247*.

Ponomareva, N.; Hazimeh, H.; Kurakin, A.; Xu, Z.; Denison, C.; McMahan, H. B.; Vassilvitskii, S.; Chien, S.; and Thakurta, A. G. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77: 1113–1201.

Protzko, J. 2018. Null-hacking, a lurking problem.

Rawal, K.; Kamar, E.; and Lakkaraju, H. 2021. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. *arxiv*.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. A. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34: 29935–29948.

Roth, A.; Tolbert, A.; and Weinstein, S. 2023. Reconciling Individual Probability Forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 101–110.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; and Zhong, C. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16: 1–85.

Rudin, C.; Zhong, C.; Semenova, L.; Seltzer, M.; Parr, R.; Liu, J.; Katta, S.; Donnelly, J.; Chen, H.; and Boner, Z. 2024. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*.

Savage, L. J. 1972. *The foundations of statistics*. Courier Corporation.

Sellam, T.; Yadlowsky, S.; Tenney, I.; Wei, J.; Saphra, N.; D'Amour, A.; Linzen, T.; Bastings, J.; Turc, I. R.; Eisenstein, J.; et al. 2022. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *International Conference on Learning Representations*.

Semenova, L. 2024. *In Pursuit of Simplicity: The Role of the Rashomon Effect for Informed Decision Making*. Ph.D. thesis, Duke University.

Semenova, L.; Chen, H.; Parr, R.; and Rudin, C. 2024. A path to simpler models starts with noise. *Advances in neural information processing systems*, 36.

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.

Shahin Shamsabadi, A.; Yaghini, M.; Dullerud, N.; Wyllie, S.; Aïvodji, U.; Alaagib, A.; Gambs, S.; and Papernot, N. 2022. Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, 35: 14170–14182.

Sharma, M.; and Bilgic, M. 2017. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery*, 31: 164–202.

Sharma, R.; Redyuk, S.; Mukherjee, S.; Sipka, A.; Vollmer, S.; and Selby, D. 2024. X Hacking: The Threat of Misguided AutoML. *arXiv preprint arXiv:2401.08513*.

Shwartz-Ziv, R.; and Armon, A. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90.

Simson, J.; Pfisterer, F.; and Kern, C. 2024. 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*, 1305–1320.

Sinha, A.; Malo, P.; and Deb, K. 2017. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE transactions on evolutionary computation*, 22(2): 276–295.

Smith, F. B.; Kossen, J.; Trollope, E.; van der Wilk, M.; Foster, A.; and Rainforth, T. 2024. Rethinking Aleatoric and Epistemic Uncertainty. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.

Sokol, K.; Kull, M.; Chan, J.; and Salim, F. 2024. Cross-model Fairness: Empirical Study of Fairness and Ethics Under Model Multiplicity. *ACM Journal on Responsible Computing*, 1(3): 1–27.

Soreanu, C.; and German, L. 2022. 5. Visual Communication in Cultural Media. The Rashomon Effect in the Image Globalization Paradigm. *Review of Artistic Education*, 26(1): 177–185.

Sreedharan, S.; Kambhampati, S.; et al. 2018. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, 518–526.

Sun, Y.; Wang, T.; and Rudin, C. 2024. Improving Decision Sparsity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tan, H. 2024. Evaluating Explanation Robustness to Model Pruning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Teney, D.; Peyrard, M.; and Abbasnejad, E. 2022. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *European Conference on Computer Vision*, 458–476. Springer.

Tindale, C. 2016. Collateral beliefs and the Rashomon effect. *Cogency*, 8(2): 125–145.

Trabasso, T. 2018. Whose memory is it? The social context of remembering. In *Memory for everyday and emotional events*, 429–443. Routledge.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33: 1633–1645.

Tran, D.; Snoek, J.; and Lakshminarayanan, B. 2020. Practical uncertainty estimation and out-of-distribution robustness in deep learning. *NeurIPS Tutorial, Google Brain*.

Tulabandhula, T.; and Rudin, C. 2014. Robust Optimization using Machine Learning for Uncertainty Sets. *ECML/PKDD 2014*, 121.

Ueki, M.; and Kawasaki, Y. 2013. Multiple choice from competing regression models under multicollinearity based on standardized update. *Computational Statistics & Data Analysis*, 63: 31–41.

Vaicenavicius, J.; Widmann, D.; Andersson, C.; Lindsten, F.; Roll, J.; and Schön, T. 2019. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, 3459–3467. PMLR.

Veran, T.; Portier, P.-E.; and Fouquet, F. 2023. Interpretable hierarchical symbolic regression for safety-critical systems with an application to highway crash prediction. *Engineering Applications of Artificial Intelligence*, 117: 105534.

Wang, A.; Kapoor, S.; Barocas, S.; and Narayanan, A. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing*, 1(1): 1–45.

Wang, A.; and Russakovsky, O. 2021. Directional bias amplification. In *International Conference on Machine Learning*, 10882–10893. PMLR.

Watson-Daniels, J. 2024. *The Roads Not Taken: Model Multiplicity in Machine Learning*. Ph.D. thesis, Harvard University Graduate School of Arts and Sciences.

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.; Calmon, F. d. P.; D'Amour, A.; Long, C.; Parkes, D. C.; and Ustun, B. 2024. Predictive churn with the set of good models. *arXiv preprint arXiv:2402.07745*.

Wu, B.; Chen, J.; Cai, D.; He, X.; and Gu, Q. 2021. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34: 7054–7067.

Wu, F.; Black, E.; and Chandrasekaran, V. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*.

Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35: 14071–14084.

Yang, Y.; and Loog, M. 2016. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2646–2651. IEEE.

Zhang, P.; Wang, H.; Naik, N.; Xiong, C.; et al. 2020. DIME: An information-theoretic difficulty measure for AI datasets. In *NeurIPS 2020 Workshop: Deep Learning through Information Geometry*.

Zhang, Y.; Khanduri, P.; Tsaknakis, I.; Yao, Y.; Hong, M.; and Liu, S. 2024. An Introduction to Bilevel Optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1): 38–59.

Zhong, C. 2024. *Interpretability and Multiplicity: a Path to Trustworthy Machine Learning*. Ph.D. thesis, Duke University.

Zhong, C.; Chen, Z.; Liu, J.; Seltzer, M.; and Rudin, C. 2024. Exploring and interacting with the set of good sparse generalized additive models. *Advances in neural information processing systems*, 36.

Zuin, G.; Marcelino, F.; Borges, L.; Couto, J.; Jorge, V.; Laurindo, M.; Barcelos, G.; Cunha, M.; Alvarenga, V.; Rodrigues, H.; et al. 2021. Predicting Heating Sliver in Duplex Stainless Steels Manufacturing through Rashomon Sets. In

2021 International Joint Conference on Neural Networks (IJCNN), 1–8. IEEE.