

# The Curious Case of Arbitrariness in Machine Learning

PRAKHAR GANESH, Mila & McGill University

AFAF TAIK, Mila

GOLNOOSH FARNADI, Mila & McGill University

Algorithmic modelling 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 traditional 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 or generalizing aspects of reality and failing to fully capture its nuances [10, 15, 25, 32, 93, 175]. It is thus inherently susceptible to arbitrariness, as it attempts to extrapolate outcomes based on limited information. Whether due to imperfect data [32, 75, 195], flawed modelling assumptions [29, 93, 101], or the unpredictability of certain tasks [59, 157, 195], this arbitrariness is an unavoidable byproduct of any data-driven learning, including machine 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 [11, 26, 29, 50, 69, 108, 162]. Recently, however, a new paradigm called *multiplicity* has emerged [20, 29, 131]. First articulated by Breiman [29], multiplicity has gained popularity due to its unique 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, commonly known as the Rashomon set. Moreover, multiplicity takes an intriguing perspective on the question of arbitrariness in model decisions, by instead examining arbitrariness in model selection. Through choices made during development and model selection, multiplicity offers an operational lens to the issue of arbitrariness and lays the groundwork for practical solutions in real-world applications.

Several existing works in the literature have provided broad overviews of the field of multiplicity. Black et al. [20] holds a special place in modern multiplicity research, offering a comprehensive discussion of “opportunities”, “concerns”, and potential “solutions” of multiplicity. Similarly, Rudin et al. [161] presents an excellent discussion on the benefits of multiplicity, with a focus on its role in identifying simpler and more interpretable models. At this point, it would be remiss not to acknowledge the PhD dissertations of Black [16], Cooper [41], Hasan [86], Hsu [95], Semenova [165], Watson-Daniels [199], Zhong [209], contributing valuable perspectives on the role of multiplicity in machine learning. Despite these notable contributions, the field still lacks a systematic review of the literature—clearly needed given its rapid growth in recent years (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.

---

Authors' Contact Information: Prakhhar Ganesh, prakhhar.ganesh@mila.quebec, Mila & McGill University; Afaf Taik, Mila, afaf.taik@mila.quebec; Golnoosh Farnadi, Mila & McGill University, farnadig@mila.quebec.

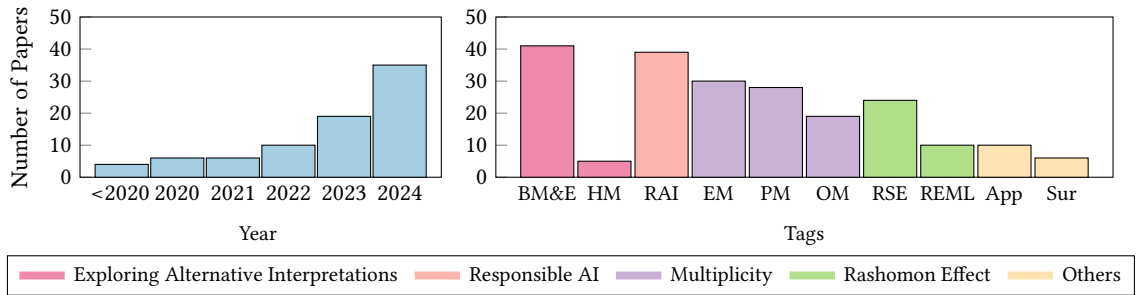


Fig. 1. Statistics from the systematic review, showing the number of papers over the years and their categorization. Each paper can have multiple tags, marking all categories of contributions made by the paper. Details of the tags are provided in the Appendix (§A). **BM&E**: Better Models and Ensembles; **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.

To ensure comprehensive coverage of the literature, 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 were eventually left with 80 papers that deeply engaged with multiplicity as a central theme in their contributions. Each paper was then manually tagged with all applicable tags, and some 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 review process are delegated to the Appendix (§A).

Building on the insights from our review, we make the following contributions to the multiplicity literature. First, we revisit the Rashomon effect, emphasizing the role of developer choices and formalizing the terminology to trace the emergence of various subdomains within multiplicity (§2). Expanding on this discussion, we extend the existing definitions of Rashomon sets and multiplicity to include underrepresented works, making both metrics of model selection and multiplicity more intentional (§3). Next, we distinguish multiplicity from related concepts such as uncertainty and variance; and provide both mathematically grounded differences as well as practical guidance on when to adopt each perspective (§4). Finally, we explore 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 and future directions to encourage continued exploration in this domain.

## 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 [5, 7, 49, 89]. Borrowing from Davis et al. [49], 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 [186, 187], the impact of culture and the fluidity of truth in ethnographic studies [89], the study of context, medium, and framing of communication [7, 181], the unreliability of eyewitnesses [92, 147], and—central to our discussion—algorithmic modelling and machine learning [20, 29].

The term Rashomon effect was first introduced into algorithmic modelling by Breiman [29], 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 [22, 30, 191, 197], null hacking [154], designing robust algorithms [34, 190], measuring variable importance [56, 66], and applications in various domains [37, 107, 182]. More recently, it has found a resurgence with increasing attention given

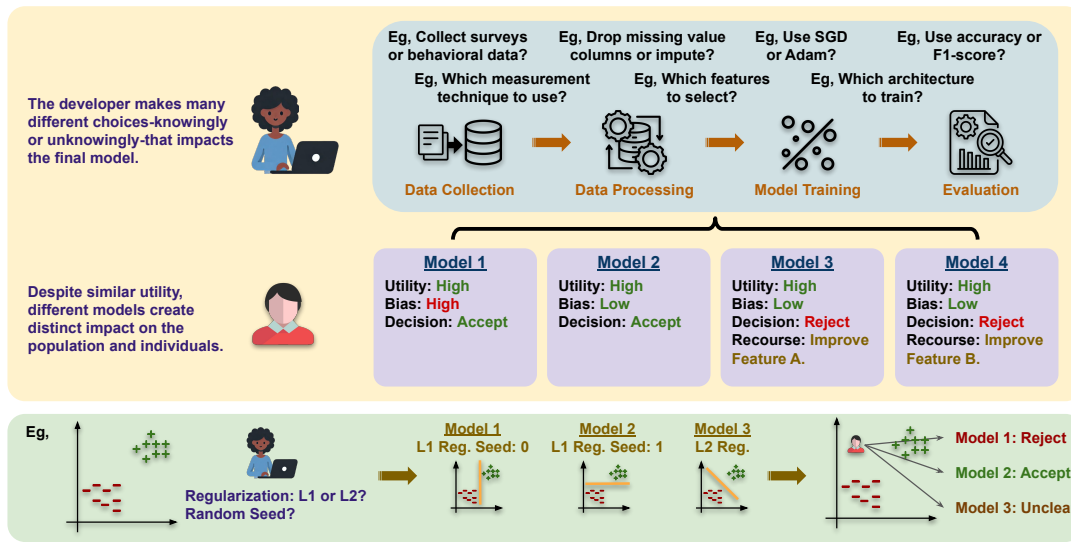


Fig. 2. Impact of developer choices on individuals and the population downstream.

to *multiplicity* in machine learning, evident both in studies that directly address the topic [13, 20, 48, 52, 131, 146, 161] and in research that situates multiplicity within the broader context of other fields [14, 105, 139, 140, 160].

**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. [207] 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”) [207], relates to how well can we 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 machine learning, resulting in gaps where multiple interpretations—and thus the Rashomon effect—can arise. While viewing the Rashomon effect through this lens of information loss is insightful, it does not provide an operational framework for developers to examine or address these challenges. Therefore, in this section, we instead focus on the role of developer choices in model design, laying the foundation for discussions on multiplicity.

## 2.1 Rashomon Effect through Design Choices and Model Selection

Designing a machine learning model involves a series of interconnected or choices. Beginning with the data, decisions are made regarding how to process and filter data, which features to select, etc. [36, 135, 174] Even the random partitioning of the data into training and validation impacts final model behavior [68]. Beyond data, the learning algorithm design further entails numerous decisions: model architecture [8, 161], hyperparameters [8, 24], various forms

of stochasticity [24, 134, 149, 151, 164], and even the evaluation and model selection criteria [72, 74]. 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 [72, 115, 153, 203]. In others, they are *conventional*, driven by popular trends or convenience [47, 61, 164, 173]. 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 the multiplicity literature, 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 certain choices as *intentional* when developers make deliberate choices with an understanding of their impact on the final model to achieve desired outcomes. Examples include incorporating bias-mitigating regularization to enhance fairness [38, 94], using simpler models for better interpretability [161], or applying data augmentation to improve robustness [64, 156]. These choices are made to steer model behaviour in a particular direction, giving the developer control over navigating the Rashomon set without the need to train multiple models, i.e., avoiding model selection. Intentional choices are typically informed by extensive prior research or other advancements that guide the developer. 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 [61, 173], such as using neural networks where simpler models would suffice [159, 173], or applying out-of-the-box fairness, robustness, or explanation techniques without understanding their implications [1, 33, 77, 124, 159, 188]. 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 [21, 47, 103, 110, 204]. As shown by Bommasani et al. [21], even shared components, for instance, 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 the model. Examples include choosing the random seed for weight initialization, data shuffling, dropout, etc., or arbitrary hyperparameter variations such as selecting between a learning rate of 0.008, 0.01, or 0.012. When dealing with arbitrary choices, making appropriate choices depends on model selection post-training. This has been widely recommended in the literature for auditing multiplicity, creating ensembles, or navigating among competing choices [47, 115, 126]. However, training multiple models can be expensive, especially for large models or complex hypothesis classes [39, 96, 98, 109, 133, 210]. Even when feasible, model selection risks overfitting, potentially undermining generalization [40, 44, 72]. 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*. For instance, consider a developer using dropout testing dropout parameter values 0.1, 0.2, and 0.3. The decision to use dropout is likely intentional, informed by a goal to enhance model robustness. However, the choice of dropout over other regularization techniques may be conventional, driven by its

widespread use or integration in popular frameworks. Finally, the specific dropout rates chosen are arbitrary, as their impact on the model’s performance is determined through selection post-training. Understanding the balance between these three factors is essential for navigating the challenges posed by the Rashomon effect.

## 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 when individuals encounter conflicting predictions from models in the Rashomon set, known as predictive multiplicity [131]. Such conflicts create brittleness and arbitrariness in decision-making, undermining the reliability of these models and hampering effective planning [20, 42, 48, 131, 137, 201]. 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) in predictions can help address the concerns of outcome homogenization [10, 47, 102, 103, 150].

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 rapidly growing field of inconsistency in explanations provided by models within the Rashomon set (37.5% of papers in our systematic review; see Figure 1). Studies have shown that models in the Rashomon set often produce conflicting variable importance and feature attribution scores [9, 19, 81, 85, 116–118, 122, 141, 145, 152, 184], which can undermine trust, for example, confusing clinicians that rely on explanations during AI-assisted diagnostics. Similarly, counterfactuals generated by one model in the Rashomon set often fail to transfer to others [87, 104, 119, 148]. 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 [119, 155].

More broadly, the effects of the Rashomon effect—multiplicity—can be measured as any form of behavioural difference between models in the context of real-world impact. 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 ~ 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 and drawing inspiration from the discussion presented here.

## 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 evaluating 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,  $\epsilon$ -Rashomon set,  $\epsilon$ -Level set, etc. We’ll stick with the term Rashomon set for consistency in our formalization. In essence, the 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 and explaining what it means for them to be indistinguishable.

Existing work in multiplicity tends to adopt a narrow view of these models. For instance, 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  $\mathcal{D}$  [34, 131, 185, 200, 202]. However, this overlooks developer choices made during data collection, data processing, and even the model training stages, all of which can influence the final model, as discussed above. Similarly, most existing studies define indistinguishability solely in terms of loss [34, 60, 83, 146, 185], disregarding other crucial choices involved in designing evaluation criteria and model selection, as previously mentioned.

To explicitly broaden the definition of Rashomon sets, we introduce a set of metric delta functions,  $\Delta^P$ , 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  $\delta_i^P \in \Delta^P$  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 3.1 (Rashomon Set).* Two models  $h_1, h_2$  belong to the same Rashomon set under performance constraints  $(\Delta^P, \mathcal{E}^P)$  if they exhibit similar performance for every metric in the given performance constraints, i.e.:

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

We now 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 [45, 73, 78, 100, 164, 196], OOD robustness multiplicity [134, 185], model complexity multiplicity [46, 125, 161, 166, 167], and feature interaction multiplicity [123], among others. To capture the various impacts discussed above, we generalize multiplicity by binding it to a metric delta  $\delta^M$  and corresponding threshold  $\epsilon^M$ , on models belonging to the Rashomon set. Formally,

*Definition 3.2 (Model Multiplicity).* Two models  $h_1, h_2$  exhibit multiplicity under performance constraints  $(\Delta^P, \mathcal{E}^P)$  and multiplicity constraint  $(\delta^M, \epsilon^M)$ , if they have similar performance for every metric in the performance constraints yet differ on the metric in the multiplicity constraint, i.e.:

$$\delta_i^P(h_1, h_2) \leq \epsilon_i^P \quad \forall (\delta_i^P, \epsilon_i^P) \in (\Delta^P, \mathcal{E}^P) \quad \text{and} \quad \delta^M(h_1, h_2) > \epsilon^M \quad (2)$$

Although our definitions are quite similar to those in 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 the model design pipeline 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

To complement the expanded definitions, 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 is particularly desirable in certain scenarios [97], because it ensures that reducing



Metric	Original Objective	Problem Setting	Monotonic	Resolution
Ambiguity [131]	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Obscurity [35]	Predictive Multiplicity	Multi-Class Classification	No	Dataset
Discrepancy [131]	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Indistinguishability [146]	Predictive Multiplicity	Multi-Class Classification	-	Dataset
Degree of Underspecification [185]	Predictive Multiplicity	Multi-Class Classification	Yes	Dataset
Viable Prediction Range [202]	Predictive Multiplicity	Probabilistic Classification	Yes	Individual
Rashomon Capacity [97]	Predictive Multiplicity	Probabilistic Classification	Yes	Individual
Multi-target Ambiguity [200]	Predictive Multiplicity	Multi-Target Classification	Yes	Dataset
Rank List Sensitivity [144]	Predictive Multiplicity	Recommender Systems	-	Dataset
Std. of Scores [90, 126]	Predictive Multiplicity	Agnostic	No	Dataset
Representational Multiplicity [90]	Procedural Multiplicity	Agnostic	No	Dataset
Region Similarity Score [180]	Procedural Multiplicity	Agnostic	-	Dataset
Variance	Arbitrariness	Agnostic	No	Individual
Self-consistency [43]	Arbitrariness	Multi-Class Classification	No	Individual
$\epsilon$ -robust to Dataset Multiplicity [135]	Dataset Multiplicity	Regression	-	Dataset
Unfairness Range [133]	Fairness Multiplicity	Agnostic	Yes	Dataset
Rashomon Ratio [167]	Size of Rashomon Set	Agnostic	Yes	Dataset
Underspecification Score [130]	Underspecification	Multi-Class Classification	No	Individual
Accuracy Under Intervention [72]	Metric Multiplicity	Multi-Class Classification	-	Dataset
Metric	Original Objective	Explanation Technique	Monotonic	Resolution
Consistency [198]	Explanation Multiplicity	Agnostic	No	Dataset
Model Class Reliance [66]	Explanation Multiplicity	Model Reliance [66]	Yes	Dataset
Attribution Agreement [31, 114]	Explanation Multiplicity	Feature Attribution	-	Individual
Profile Disparity Index [111]	Explanation Multiplicity	Model Profile [111]	-	Dataset
Inv. Cost of Neg. Surprise [148]	Explanation Multiplicity	Counterfactuals	-	Dataset
Variable Importance Clouds [56]	Explanation Multiplicity	Feature Attribution	Yes	Dataset
Coverage & Interval Width [132]	Explanation Uncertainty	Agnostic	No	Individual

Table 1. Existing multiplicity metrics in the literature, along with original objective, problem setting, monotonicity wrt to the Rashomon set, and metric resolution. Metrics with no entry under ‘Monotonic’ are defined to compare two models, and not the Rashomon set.

the Rashomon set size—by improving specification—will either decrease or maintain the multiplicity. However, in other contexts, metrics that are not necessarily monotonic, such as variance, can be equally valuable. For instance, when it is important to quantify not just the range but also the dispersion of arbitrariness, metrics that capture such variability become crucial. Ultimately, the relevance of monotonicity in a metric depends on the specific context in which it is applied. 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 in evaluating multiplicity.

Beyond the metrics themselves, we also revisit the challenge of training multiple models to evaluate multiplicity, exploring more efficient alternatives. Madras et al. [130] propose a local ensembling technique to quantify underspecification by analyzing the loss curvature, eliminating the need to train multiple models. Hsu et al. [98] 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 [109] 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 [96, 133, 205, 210]. 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 [69, 70, 108, 177]—model’s lack of confidence in its predictions—or the bias-variance decomposition [28, 55, 112]—dividing the model error into how well the model fits the data (bias) and its sensitivity to changes in data (variance). With extensive literature already present in these areas, a natural question arises: *What unique perspectives does multiplicity bring to the discussion of arbitrariness not already covered by these concepts?* In this section, we formalize the interplay between multiplicity, uncertainty, and bias-variance decomposition, addressing this question mathematically and through practical recommendations for when different perspectives on arbitrariness are most valuable.

### 4.1 Multiplicity and Uncertainty

**Prediction Uncertainty:** We start by defining uncertainty, drawing heavily from Gal [69], Gal et al. [70], Kendall and Gal [108], Smith et al. [177]. 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:

$$U(x, D) = H_y[\text{Prob}(y|x, D)] = H_y[\Sigma_\theta[\text{Prob}(y|x, \theta) * \text{Prob}(\theta|D)]] \quad (3)$$

$$= H_y[E_{\theta \sim \text{Prob}(\theta|D)}[\text{Prob}(y|x, \theta)]] \quad (4)$$

where  $x$  is the input for which we’re measuring uncertainty,  $y$  is the output variable,  $D$  is the original training data, and  $\theta$  is the parametric representation of learned models.  $U(x, D)$  is the prediction uncertainty, while  $H[\cdot]$ ,  $E[\cdot]$ ,  $\text{Prob}[\cdot]$  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.

Prediction uncertainty is also typically divided into aleatoric and epistemic uncertainty [99, 108, 168]. Aleatoric uncertainty, connected to “distributional complexity” [207], is the intrinsic difficulty of modelling the relationship in the data distribution, often described as the uncertainty left with access to infinite data. In contrast, epistemic uncertainty, connected to “approximation complexity” [207], stems from a lack of knowledge about the best model to use, i.e., the challenge of accurately capturing the true data distribution. Epistemic uncertainty does not have a straightforward formulation. Instead, it is commonly defined as the uncertainty remaining after accounting for aleatoric uncertainty.

For the interested reader, we encourage exploring the uncertainty literature [69, 99, 108, 189], as we do not expand on these concepts in our work. We simply restate these definitions to compare them directly 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:

$$M(x, D) = H_y[E_{\theta \sim \text{Prob}_R(\theta|D)}[\text{Prob}(y|x, \theta)]] \quad \text{s.t.} \quad \text{Prob}_R(\theta|D) = \begin{cases} 0 & \text{if } \theta \notin R \\ \frac{\text{Prob}(\theta|D)}{\sum_{\theta \in R} \text{Prob}(\theta|D)} & \text{if } \theta \in R \end{cases} \quad (5)$$

where  $\text{Prob}_R(\theta|D)$  represents 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). 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 relationship between data and the predictions derived from it. Multiplicity, on the other hand, offers the most relevant perspective for actively exploring 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 whether to use the lens of multiplicity or uncertainty.

- *Uncertainty provides an information-centric perspective.* As uncertainty definitions are derived from information theory [26, 69, 108], 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’ [142, 171, 206].
- *Uncertainty is sufficient when dealing with distributional complexity.* Noise in real-world data may result in a lack of predictive power to make reliable decisions [195]. Having access to different interpretations through multiplicity adds little value in such cases—multiple ways of delivering incorrect answers do not necessarily enhance usefulness.
- *Uncertainty quantification can be more efficient, but research in multiplicity quantification is growing rapidly.* Uncertainty is streamlined into modern machine learning pipelines through Bayesian networks and model calibration [106, 143, 192], offering a cost-effective alternative to multiplicity. Even when training multiple models, there is typically no definitive way to ensure that every trained model falls within the Rashomon set, and thus not all trained models contribute to measuring multiplicity whereas they are still valuable for quantifying uncertainty. That said, advancements in multiplicity research, as previously discussed in §3.2, have already been challenging this dynamic and may continue to reshape it further in the future [96, 98, 109, 130, 133, 210].
- *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. [36, 72, 73, 100, 135], or, broadly, any form of bi-level or constrained optimization, including meta-objectives [18, 73, 76, 166, 183].
- *Multiplicity is better suited to explore alternative interpretations.* As we’ve discussed, 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. [19, 21, 29, 47, 118, 126, 129, 204]

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 Bias-Variance Decomposition

Another common measure of arbitrariness in decisions is the ‘variance’ from the bias-variance decomposition. Error in machine learning is often categorized into three parts: irreducible error, bias, and variance [28, 55, 112]. This is known as the bias-variance decomposition. 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.

Understandably, at first glance, sensitivity to variations in the dataset—captured by ‘variance’—might seem like a natural way to measure arbitrariness in decision-making. However, bias-variance decomposition is typically confined to the analysis of a single model, focusing only on sensitivity to variations in the underlying dataset. While this framework provides valuable insights into a model’s behaviour—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,

$$Error = IrreducibleError + Bias + \text{Var}_{\mathcal{D}}[f_{\theta}(\mathbf{x})] \quad [28, 55, 112] \quad (6)$$

$$PredictiveMultiplicity = \text{Var}_{\theta \sim \text{Prob}_{\mathcal{R}}(\theta|\mathcal{D})}[f_{\theta}(\mathbf{x})] \quad [83, 90, 126] \quad (7)$$

where  $f_{\theta}(\cdot)$  is the predictive function for the learned parameter value  $\theta$ , and  $\text{Var}[\cdot]$  represents variance. We use a metric of predictive multiplicity that also uses variance to quantify multiplicity, facilitating clearer comparisons [83, 90, 126]. 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. As we will explore further in §5, multiplicity allows us to navigate through multiple candidate models and even identify those with desirable bias-variance properties—such as ‘simpler’ models, often associated with lower variance [46, 125, 161, 166, 167].

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. It is also crucial to recognize the common approximations made in the literature and to explicitly acknowledge their assumptions. For instance, predictive multiplicity has been used in the literature to measure epidemic uncertainty [4, 31, 169]. This stems from the intuition that epistemic uncertainty is often estimated using Bayesian neural networks, whose behaviour can resemble Rashomon sets. However, this approximation is only valid if the models are calibrated, meaning they already account for aleatoric uncertainty [99]. Furthermore, it is important to distinguish between multiplicity within the Rashomon set and uncertainty (of any kind) across the entire set of models, as discussed earlier. Thus, using multiplicity as a proxy for epistemic uncertainty is an approximation which may fail in situations with high aleatoric uncertainty or when the distribution over the Rashomon set is significantly different from that on the entire set of models.

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 with Multiplicity

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 uncover and explore various ‘good’ learned interpretations [20, 161]. 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. Additionally, moving beyond a single-model paradigm also opens the door to aggregating insights from multiple models. In this section, we study how multiplicity facilitates such exploration and its broader implications. Our discussion builds on what Black et al. [20] 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 commonly seen 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 [176, 208]. A classic example is hyperparameter optimization. Constrained optimization, on the other hand, involves solving the optimization problems under specific constraints [12, 71, 80]. Examples include enforcing constraints of fairness, robustness, safety, etc. during

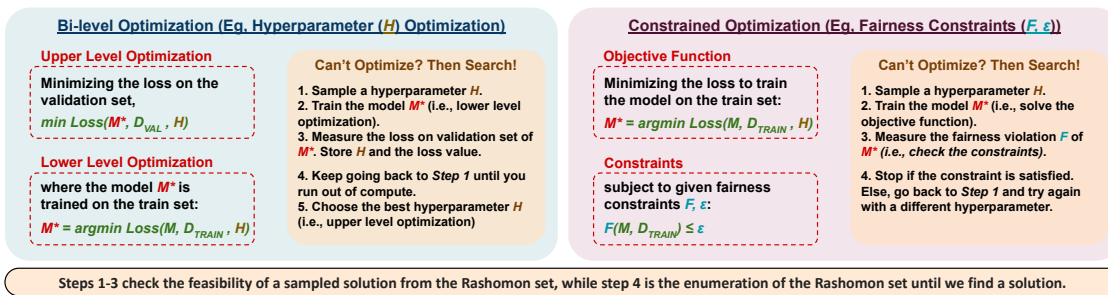


Fig. 3. The role of multiplicity in brute-force search for bi-level and constrained optimization problems in machine learning.

model training [62, 71]. These classes of optimization problems can be notoriously difficult to solve. Challenges arise from the complexity of formalizing certain constraints, the difficulty of creating continuous and 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 [71, 176, 208]. 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 3). Multiplicity has been used in the literature to find models with lower bias [18, 74, 100, 174], smaller model complexity [23, 46, 54, 161, 166, 167], better explanations [183, 193], improved generalizability [122], and the ability to allow personalization [129]. 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 [63, 88, 128].

## 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 can often yield better results in such scenarios. Techniques like prediction ensembling, or bagging, have long been a central recommendation for stability in machine learning [27, 53, 126]. Many methods in the multiplicity literature have capitalized on combining various forms of information from models in the Rashomon set. While literature in this direction primarily focuses on aggregating model explanations to create more stable and reliable explanations [9, 19, 57, 58, 66, 81, 82, 91, 104, 111, 113, 118, 121, 123, 132, 138, 178, 182, 211], other works have also shown the benefits of aggregating fairness scores [45], individual probabilities [158], or regression analysis to discover causality [191]. 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 enable the discovery of better models, it also introduces risks, particularly the potential of exploiting these search methods to circumvent regulatory requirements and interventions. This is more prevalent when broad principles, such as fairness, are reduced to specific benchmarks or metrics [17]. By leveraging multiplicity, it becomes possible to ‘hack’ these metrics, producing models that meet the specified criteria without truly adhering to the underlying principles the metrics are intended to enforce [40, 44, 139]. Several studies in the literature have

shown that such a search can indeed produce models capable of “regulatory-washing”, being able to manipulate explanations [2, 170, 172] and fairness scores [17, 67, 74]. Such manipulation can also occur unintentionally—a result of overfitting to a given metric [17, 44, 72]—underscoring the need for vigilance against the misuse of multiplicity and calling for a more robust operationalization of regulatory frameworks [17].

## 6 Multiplicity and Responsible AI

We now move to placing 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 et al. [20], 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, particularly in critical domains such as law and medicine, where they can have direct and lasting impacts on human lives [10, 20, 48, 79, 179, 201]. Borrowing an analogy from Gomez et al. [79], imagine a judge deciding legal cases by flipping a coin. While this may seem extreme, it demonstrates how machine learning 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 by the developers 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 where individuals lack access to other ‘*equivalent opportunities*’ [10]. For instance, in hiring, some level of arbitrariness may be acceptable, or even useful to deal with the concerns of homogenization (§6.2). This is because individuals looking for a job often seek multiple opportunities across different 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 critical situations, the presence of arbitrariness raises serious concerns for the responsible deployment of machine learning models.

As explored earlier, the negative effects of multiplicity extend beyond the arbitrariness of just the final prediction or decision (§2.2). For instance, multiplicity in counterfactual explanations can impact the stability and validity of algorithmic recourse [9, 84, 85, 87, 104, 111, 119, 120, 182]. 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 option provided by one model may become invalid if the model is updated, invalidating previous efforts. This inconsistency raises both ethical and legal concerns [51, 111].

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. In modern machine learning, 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 [10, 15, 32, 75, 93]. Research has consistently shown that such disparities exacerbate existing inequalities—whether through arbitrariness, uncertainty, or multiplicity [65]. The resulting disproportionate harms across groups highlight the pressing need to address arbitrariness in critical domains [4, 43, 79].

**Impact of Other Responsible AI Constraints on 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

Term	Definition
Algorithmic Blackballing [3]	'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 [110]	'The notion that choices and preferences will become homogeneous in the face of algorithmic curation'
Algorithmic Leviathan [47]	'Automated decision-making systems that make uniform judgments across broad swathes of a sector.'
Outcome Homogenization [21]	'The phenomenon of individuals (or groups) exclusively receiving negative outcomes from all ML models they interact with'
Generative Monoculture [204]	'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 [103]	'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.

responsible AI. For instance, studies have shown that imposing fairness constraints can inadvertently increase multiplicity [35, 126]. However, Long et al. [126] 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. [115] 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 [47] 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 [110] 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. [21] 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*.

Interestingly, introducing controlled arbitrariness or multiplicity can mitigate these risks by preventing monocultures. In contexts where arbitrariness in individual decisions is less concerning but homogenization is, controlled multiplicity becomes a desirable property [10, 102, 150]. 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 [10, 102, 150], public perception of intentional randomness in decision-making remains skeptical. A recent study by Meyer et al. [136] 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 [10].

## 7 Open Research Questions and Emerging Trends

Building upon our systematic survey of the literature on multiplicity, this work gives us a unique vantage point to identify and discuss several emerging trends and critical open research questions in the field,

- *Expanding the scope of multiplicity beyond predictions and explanations.* One of our key motivations 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 deeper and 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 in machine learning 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.* While substantial work has been done on measuring and mitigating multiplicity, many studies do not engage further with the concerns of multiplicity in real-world settings. As discussed, arbitrariness can be both a force of good (§6.2) or a cause of harm (§6.1), depending on the context. 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 crucial to truly grasp the real-world consequences of multiplicity.
- *Multiplicity in the era of LLMs.* As machine learning models continue to scale, particularly in the era of large language models (LLMs), 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 another dimension of multiplicity in LLMs—decisions multiplicity due to changing prompts—that requires deeper examination [6, 127, 163, 194].

## 8 Conclusion

In this work, we systemized the existing knowledge on multiplicity, uncovering interesting trends and insights. 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.



## Acknowledgments

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google 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

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).
- [2] Ulrich Aivodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2021. Characterizing the risk of fairwashing. *Advances in Neural Information Processing Systems* 34 (2021), 14822–14834.
- [3] Ifeoma Ajunwa. 2020. An auditing imperative for automated hiring systems. *Harv. JL & Tech.* 34 (2020), 621.
- [4] Junaid Ali, Preethi Lahoti, and Krishna P Gummadi. 2021. Accounting for model uncertainty in algorithmic discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 336–345.
- [5] Valerie Alia. 1996. The Rashomon principle: The journalist as ethnographer. *Deadlines and Diversity: Journalism Ethics in a Changing World*. Halifax: Fernwood Publishing (1996).
- [6] Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwaresh, Areeb Alowisheq, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781* (2024).
- [7] Robert Anderson. 2016. The Rashomon effect and communication. *Canadian Journal of Communication* 41, 2 (2016), 249–270.
- [8] Christian Arnold, Luka Biedebach, Andreas Küpfer, and Marcel Neunhoeffer. 2024. The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods* 12, 4 (2024), 841–848.
- [9] Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. 2024. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery* 38, 5 (2024), 2596–2632.
- [10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- [11] Thomas Bayes. 1763. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions* 53 (1763), 370–418. Published posthumously, communicated by Richard Price.
- [12] Dimitri P Bertsekas. 2014. *Constrained optimization and Lagrange multiplier methods*. Academic press.
- [13] Przemyslaw Biecek, Hubert Baniecki, Mateusz Krzyżiński, and Dianne Cook. 2024. Performance is not enough: the story told by a Rashomon quartet. *Journal of Computational and Graphical Statistics* just-accepted (2024), 1–6.
- [14] Przemyslaw Biecek and Wojciech Samek. 2024. Position: Explain to Question not to Justify. In *Forty-first International Conference on Machine Learning*.
- [15] Abeba Birhane. 2022. *Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence*. Ph.D. Dissertation. University College Dublin. School of Computer Science.
- [16] Emily Black. 2022. *(Un) Fairness Along the AI Pipeline Problems and Solutions*. Ph.D. Dissertation. Carnegie Mellon University.
- [17] Emily Black, Talia Gillis, and Zara Yasmine Hall. 2024. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 602–615.
- [18] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. Less discriminatory algorithms. *Georgetown Law Journal* 113, 1 (2024).
- [19] Emily Black, Klas Leino, and Matt Fredrikson. 2021. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.
- [20] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 850–863.
- [21] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.
- [22] Peter L Bonate. 2006. *The art of modeling*. Springer.
- [23] Zachery Boner, Harry Chen, Lesia Semenova, Ronald Parr, and Cynthia Rudin. [n. d.]. Using Noise to Infer Aspects of Simplicity Without Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [24] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems* 3 (2021), 747–769.
- [25] George EP Box. 1976. Science and statistics. *J. Amer. Statist. Assoc.* 71, 356 (1976), 791–799.
- [26] G. E. P. Box and George C. Tiao. 1973. Bayesian inference in statistical analysis. *International Statistical Review* 43 (1973), 242. <https://api.semanticscholar.org/CorpusID:122028907>



- [27] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24 (1996), 123–140.
- [28] L BREIMAN. 1996. Bias, variance, and arcing classifiers. *Univ. of California, Dept. of Stat., TR 460* (1996).
- [29] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.
- [30] David I Broadhurst and Douglas B Kell. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2 (2006), 171–196.
- [31] Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. 2022. Implications of model indeterminacy for explanations of automated decisions. *Advances in Neural Information Processing Systems* 35 (2022), 7810–7823.
- [32] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [33] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705* (2019).
- [34] Enrique Castillo, Roberto Mínguez, Carmen Castillo, and Antonio S Cofino. 2008. Dealing with the multiplicity of solutions of the l1 and l1-inf regression models. *European journal of operational research* 188, 2 (2008), 460–484.
- [35] Mustafa Cavus and Przemyslaw Biecek. 2024. An Experimental Study on the Rashomon Effect of Balancing Methods in Imbalanced Classification. *arXiv preprint arXiv:2405.01557* (2024).
- [36] Mustafa Cavus and Przemyslaw Biecek. 2024. Investigating the Impact of Balancing, Filtering, and Complexity on Predictive Multiplicity: A Data-Centric Perspective. *arXiv preprint arXiv:2412.09712* (2024).
- [37] Guillermo R Chantre, Mario R Vigna, Juan P Renzi, and Anibal M Blanco. 2018. A flexible and practical approach for real-time weed emergence prediction based on Artificial Neural Networks. *Biosystems engineering* 170 (2018), 51–60.
- [38] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–30.
- [39] Martino Ciaperoni, Han Xiao, and Aristides Gionis. 2024. Efficient exploration of the Rashomon set of rule-set models. In *Proceedings Of The 30th ACM SIGKDD Conference On Knowledge Discovery And Data Mining*. 478–489.
- [40] Beau Coker, Cynthia Rudin, and Gary King. 2021. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science* 67, 10 (2021), 6174–6197.
- [41] A Feder Cooper. 2024. *Between Randomness and Arbitrariness: Some Lessons for Reliable Machine Learning at Scale*. Ph.D. Dissertation. Cornell University.
- [42] A Feder Cooper, Jonathan Frankle, and Christopher De Sa. 2022. Non-Determinism and the Lawlessness of Machine Learning Code. In *Proceedings of the 2022 Symposium on Computer Science and Law*. 1–8.
- [43] A Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22004–22012.
- [44] A Feder Cooper, Yucheng Lu, Jessica Forde, and Christopher M De Sa. 2021. Hyperparameter optimization is deceiving us, and how to stop it. *Advances in Neural Information Processing Systems* 34 (2021), 3081–3095.
- [45] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*. PMLR, 2144–2155.
- [46] Evzenie Coupkova and Mireille Boutin. 2024. On the Rashomon ratio of infinite hypothesis sets. *arXiv preprint arXiv:2404.17746* (2024).
- [47] Kathleen Creel and Deborah Hellman. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43.
- [48] Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23, 226 (2022), 1–61.
- [49] Blair Davis, Robert Anderson, and Jan Walls. 2015. *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.
- [50] Abraham de Moivre. 1718. *The Doctrine of Chances: Or, A Method of Calculating the Probabilities of Events in Play*. W. Pearson, London. First edition.
- [51] Giovanni De Toni, Stefano Teso, Bruno Lepri, and Andrea Passerini. 2024. Time Can Invalidate Algorithmic Recourse. *arXiv preprint arXiv:2410.08007* (2024).
- [52] Marco Del Giudice. 2024. The prediction-explanation fallacy: a pervasive problem in scientific applications of machine learning. *Methodology* 20, 1 (2024), 22–46.
- [53] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.
- [54] Katarina Doctor, Tong Mao, and Hrushikesh Mhaskar. 2024. Encoding of data sets and algorithms. *Applied Numerical Mathematics* 200 (2024), 209–235.
- [55] Pedro Domingos. 2000. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*. Morgan Kaufmann Stanford, 231–238.
- [56] Jiayun Dong and Cynthia Rudin. 2019. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209* (2019).

- [57] Jiayun Dong and Cynthia Rudin. 2020. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence* 2, 12 (2020), 810–824.
- [58] Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward Browne. 2023. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Advances in Neural Information Processing Systems* 36 (2023), 6267–6279.
- [59] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaa05580.
- [60] Ally Yalei Du, Dung Daniel Ngo, and Zhiwei Steven Wu. 2024. Reconciling Model Multiplicity for Downstream Decision Making. *arXiv preprint arXiv:2405.19667* (2024).
- [61] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. 2022. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *Neurocomputing* 503 (Sept. 2022), 92–108. <https://doi.org/10.1016/j.neucom.2022.06.111>
- [62] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
- [63] Gilles Eerlings, Sebe Vanbrabant, Jori Liesenborgs, Gustavo Rovelo Ruiz, Davy Vanackén, and Kris Luyten. 2024. AI-Spectra: A Visual Dashboard for Model Multiplicity to Enhance Informed and Transparent Decision-Making. *arXiv preprint arXiv:2411.10490* (2024).
- [64] Hamid Eghbal-zadeh, Werner Zellinger, Maura Pintor, Kathrin Grosse, Khaled Koutini, Bernhard A Moser, Battista Biggio, and Gerhard Widmer. 2024. Rethinking data augmentation for adversarial robustness. *Information Sciences* 654 (2024), 119838.
- [65] Golnoosh Farnadi, Mohammad Havaei, and Negar Rostamzadeh. 2024. Position: Cracking the Code of Cascading Disparity Towards Marginalized Communities. In *Forty-first International Conference on Machine Learning*.
- [66] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 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 (2019), 1–81.
- [67] Jessica Zosa Forde, A Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman. 2021. Model selection’s disparate impact in real-world deep learning applications. *arXiv preprint arXiv:2104.00606* (2021).
- [68] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.
- [69] Yarin Gal. 2016. *Uncertainty in deep learning*. Ph. D. Dissertation. University of Cambridge.
- [70] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*. PMLR, 1183–1192.
- [71] Jose Gallego-Posada. 2024. *Constrained optimization for machine learning: algorithms and applications*. Ph. D. Dissertation. Université de Montréal.
- [72] Prakhar Ganesh. 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.
- [73] Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. 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.
- [74] Prakhar Ganesh, Usman Gohar, Lu Cheng, and Golnoosh Farnadi. 2024. Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML. *arXiv preprint arXiv:2411.11101* (2024).
- [75] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
- [76] Marissa Gerchick, Tobi Jegede, Tarak Shah, Ana Gutierrez, Sophie Beiers, Noam Shemtov, Kath Xu, Anjana Samant, and Aaron Horowitz. 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.
- [77] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. 2023. AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology* 96, 1150 (2023), 20230023.
- [78] Talia B Gillis, Vitaly Meursault, and Berk Ustun. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 377–387.
- [79] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. 2024. Algorithmic Arbitrariness in Content Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2234–2253.
- [80] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [81] Abirami Gunasekaran, Pritesh Mistry, and Minsi Chen. 2024. Which Explanation Should be Selected: A Method Agnostic Model Class Reliance Explanation for Model and Explanation Multiplicity. *SN Computer Science* 5, 5 (2024), 1–20.
- [82] Masaki Hamamoto and Masashi Egi. 2021. Model-agnostic ensemble-based explanation correction leveraging rashomon effect. In *2021 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 01–08.
- [83] Faisal Hamman, Pasan Dissanayake, Saumitra Mishra, Freddy Lecue, and Sanghamitra Dutta. 2024. Quantifying prediction consistency under model multiplicity in tabular LLMs. *arXiv preprint arXiv:2407.04173* (2024).
- [84] Faisal Hamman, Erfaan Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2023. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning*. PMLR, 12351–12367.

- [85] Leif Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 640–647.
- [86] Md Golam Moula Mehedi Hasan. 2022. *Targeted Generation of Counterfactual Examples for Data Augmentation and to Mitigate the Rashomon Effect*. Ph. D. Dissertation. Tennessee Technological University.
- [87] Md Golam Moula Mehedi Hasan and Douglas Talbert. 2022. Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach. In *The International FLAIRS Conference Proceedings*, Vol. 35.
- [88] Zezhen He and Yaron Shaposhnik. 2023. Visualizing the Implicit Model Selection Tradeoff. *Journal of Artificial Intelligence Research* 76 (2023), 829–881.
- [89] Karl G Heider. 1988. The Rashomon effect: When ethnographers disagree. *American Anthropologist* 90, 1 (1988), 73–81.
- [90] Ari Heljakka, Martin Trapp, Juho Kannala, and Arno Solin. 2022. Disentangling model multiplicity in deep learning. *arXiv preprint arXiv:2206.08890* (2022).
- [91] Martin Thomas Hibbeln, Raphael M. Kopp, and Noah Urban. 2024. Heterogeneous Machine Learning Ensembles for Recovery Rate Forecasting. (2024).
- [92] William Hirst and Adam Brown. 2011. On the Virtues of an Unreliable Memory. *Grounding sociality: Neurons, mind, and culture* (2011), 95–113.
- [93] Sara Hooker. 2021. Moving beyond “algorithmic bias is a data problem”. *Patterns* 2, 4 (2021).
- [94] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 1–52.
- [95] Hsiang Hsu. 2023. *Information-Theoretic Tools for Machine Learning Beyond Accuracy*. Ph. D. Dissertation. Harvard University Graduate School of Arts and Sciences.
- [96] Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Chun-Fu Chen. [n. d.]. RashomonGB: Analyzing the Rashomon Effect and Mitigating Predictive Multiplicity in Gradient Boosting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [97] Hsiang Hsu and Flavio Calmon. 2022. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems* 35 (2022), 28988–29000.
- [98] Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. 2024. Dropout-Based Rashomon Set Exploration for Efficient Predictive Multiplicity Estimation. In *The Twelfth International Conference on Learning Representations*.
- [99] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110, 3 (2021), 457–506.
- [100] Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can we obtain fairness for free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 586–596.
- [101] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.
- [102] Shomik Jain, Kathleen Creel, and Ashia Wilson. 2024. Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. *arXiv preprint arXiv:2404.08592* (2024).
- [103] Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 197–206.
- [104] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Recourse under model multiplicity via argumentative ensembling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 954–963.
- [105] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2024. Robust Counterfactual Explanations in Machine Learning: A Survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 8086–8094. Survey Track.
- [106] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bannamoun. 2022. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine* 17, 2 (2022), 29–48.
- [107] John Kang, Tiziana Rancati, Sangkyu Lee, Jung Hun Oh, Sarah L Kerns, Jacob G Scott, Russell Schwartz, Seyoung Kim, and Barry S Rosenstein. 2018. Machine learning and radiogenomics: lessons learned and future directions. *Frontiers in oncology* 8 (2018), 228.
- [108] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).
- [109] Nicholas Kissel and Lucas Mentch. 2024. Forward stability and model path selection. *Statistics and Computing* 34, 2 (2024), 82.
- [110] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.
- [111] Katarzyna Kobylińska, Mateusz Krzyżiński, Rafał Machowicz, Mariusz Adamek, and Przemysław Biecek. 2024. Exploration of the Rashomon Set Assists Trustworthy Explanations for Medical Data. *IEEE Journal of Biomedical and Health Informatics* 28, 11 (2024), 6454–6465.
- [112] Eun Bae Kong and Thomas G. Dietterich. 1995. Error-correcting output coding corrects bias and variance. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML’95)*. Morgan Kaufmann Publishers Inc., 313–321.
- [113] Daniel R Kowal. 2022. Bayesian subset selection and variable importance for interpretable prediction and classification. *Journal of Machine Learning Research* 23, 108 (2022), 1–38.

- [114] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner’s perspective. *arXiv preprint arXiv:2202.01602* (2022).
- [115] Bogdan Kulynych, Hsiang Hsu, Carmela Troncoso, and Flavio P Calmon. 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.
- [116] Rishabh Kumar, Adriano Koshiyama, Kleyton da Costa, Nigel Kingsman, Marvin Tewarrie, Emre Kazim, Arunita Roy, Philip Treleaven, and Zac Lovell. 2023. *Deep learning model fragility and implications for financial stability and regulation*. Technical Report. Bank of England.
- [117] Jakub Kuzilek and Mustafa Çavuş. 2024. Rashomon effect in Educational Research: Why More is Better Than One for Measuring the Importance of the Variables? *arXiv preprint arXiv:2412.12115* (2024).
- [118] Gabriel Laberge, Yann Pequinot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. 2023. Partial order in chaos: consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research* 24, 364 (2023), 1–50.
- [119] Francesco Leofante, Elena Botoeva, and Vineet Rajani. 2023. Counterfactual explanations and model multiplicity: a relational verification view. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, Vol. 19. 763–768.
- [120] Anastasia-M Leventi-Peetz and Kai Weber. 2022. Rashomon effect and consistency in explainable artificial intelligence (XAI). In *Proceedings of the Future Technologies Conference*. Springer, 796–808.
- [121] Dan Ley, Leonard Tang, Matthew Nazari, Hongjin Lin, Suraj Srinivas, and Himabindu Lakkaraju. 2023. Consistent Explanations in the Face of Model Indeterminacy via Ensembling. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.
- [122] Sichao Li, Amanda S Barnard, and Quanling Deng. 2024. Practical attribution guidance for rashomon sets. *arXiv preprint arXiv:2407.18482* (2024).
- [123] Sichao Li, Rong Wang, Quanling Deng, and Amanda S Barnard. [n. d.]. Exploring the cloud of feature interaction scores in a Rashomon set. In *The Twelfth International Conference on Learning Representations*.
- [124] Zachary C Lipton. 2018. The myths of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
- [125] Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, and Cynthia Rudin. 2022. FasterRisk: Fast and accurate interpretable risk scores. *Advances in Neural Information Processing Systems* 35 (2022), 17760–17773.
- [126] Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. 2024. Individual arbitrariness and group fairness. *Advances in Neural Information Processing Systems* 36 (2024).
- [127] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8086–8098.
- [128] Kris Luyten, Gilles Eerlings, Jori Liesenborgs, Gustavo Rovelto Ruiz, Sebe Vanbrabant, and Davy Vanackem. 2024. Opportunities and Challenges of Model Multiplicity in Interactive Software Systems. (2024).
- [129] Sisi Ma, Pamela J Schreiner, Elizabeth R Seaquist, Mehmet Ugurbil, Rachel Zmora, and Lisa S Chow. 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 (2020), 103379.
- [130] David Madras, James Atwood, and Alexander D’Amour. 2019. Detecting extrapolation with local ensembles. In *International Conference on Learning Representations*.
- [131] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.
- [132] Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. 2023. But are you sure? an uncertainty-aware perspective on explainable ai. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 7375–7391.
- [133] Kota Mata, Kentaro Kanamori, and Hiroki Arimura. 2022. Computing the Collection of Good Models for Rule Lists. In *Proc. the 18th International Conference on Machine Learning and Data Mining (MLDM 2022)*.
- [134] R Thomas McCoy, Junghyun Min, and Tal Linzen. 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.
- [135] Anna P Meyer, Aws Albarghouthi, and Loris D’Antoni. 2023. The dataset multiplicity problem: How unreliable data impacts predictions. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 193–204.
- [136] Anna P Meyer, Yea-Seul Kim, Aws Albarghouthi, and Loris D’Antoni. 2024. Perceptions of the Fairness Impacts of Multiplicity in Machine Learning. *arXiv preprint arXiv:2409.12332* (2024).
- [137] Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. 2016. Launch and iterate: Reducing prediction churn. *Advances in Neural Information Processing Systems* 29 (2016).
- [138] Oana Mitruț, Gabriela Moise, Alin Moldoveanu, Florica Moldoveanu, Marius Leordeanu, and Livia Petrescu. 2024. Clarity in complexity: how aggregating explanations resolves the disagreement problem. *Artificial Intelligence Review* 57, 12 (2024), 1–53.
- [139] Kajsa Møllersen and Einar Holsbø. 2023. Accounting for multiplicity in machine learning benchmark performance. *arXiv preprint arXiv:2303.07272* (2023).
- [140] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International*

- Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 39–68.
- [141] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. 2023. An empirical evaluation of the Rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 462–478.
- [142] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning* 111, 1 (2022), 89–122.
- [143] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.
- [144] Sejoon Oh, Berk Ustun, Julian McAuley, and Srijan Kumar. 2022. Rank list sensitivity of recommender systems to interaction perturbations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1584–1594.
- [145] Kotaro Okazaki, Tony Ribeiro, Kuo-Yen Lo, Jyunichi Sakuma, and Katsumi Inoue. 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). Japan Society for Artificial Intelligence, 2Q5IS104–2Q5IS104.
- [146] Lucas Monteiro Paes, Rodrigo Cruz, Flavio P Calmon, and Mario Diaz. 2023. On the inevitability of the Rashomon effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 549–554.
- [147] Ainat Pansky, Asher Koriati, and Morris Goldsmith. 2005. Eyewitness recall and testimony. *Psychology and law: An empirical perspective* (2005), 93–150.
- [148] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 809–818.
- [149] Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. A survey on stability of learning with limited labelled data and its sensitivity to the effects of randomness. *Comput. Surveys* 57, 1 (2024), 1–40.
- [150] Ronen Perry and Tal Zarsky. 2015. 'May the Odds Be Ever in Your Favor': Lotteries in Law. *Alabama Law Review* 66 (2015), 1035–1098.
- [151] David Picard. 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* (2021).
- [152] Clement Poiret, Antoine Grigis, Justin Thomas, and Marion Noulhiane. 2023. Can we Agree? On the Rashomon Effect and the Reliability of Post-Hoc Explainable AI. *arXiv preprint arXiv:2308.07247* (2023).
- [153] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research* 77 (2023), 1113–1201.
- [154] John Protzko. 2018. Null-hacking, a lurking problem. (2018).
- [155] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 2021. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. *arxiv* (2021).
- [156] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems* 34 (2021), 29935–29948.
- [157] Nicholas Rescher. 1997. *Predicting the future: An introduction to the theory of forecasting*. State University of New York Press.
- [158] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2023. Reconciling Individual Probability Forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 101–110.
- [159] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
- [160] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 16 (2022), 1–85.
- [161] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. 2024. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846* (2024).
- [162] Leonard J Savage. 1972. *The foundations of statistics*. Courier Corporation.
- [163] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.
- [164] Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, et al. [n. d.]. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *International Conference on Learning Representations*.
- [165] Lesia Semenova. 2024. *In Pursuit of Simplicity: The Role of the Rashomon Effect for Informed Decision Making*. Ph.D. Dissertation. Duke University.
- [166] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. 2024. A path to simpler models starts with noise. *Advances in neural information processing systems* 36 (2024).
- [167] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1827–1858.
- [168] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. 2014. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences* 255 (2014), 16–29.

- [169] RM Shahbab and Kais Zaman. 2024. A new approach to robustness-based optimization using uncertainty set constructed through machine learning. *Structural and Multidisciplinary Optimization* 67, 7 (2024), 125.
- [170] Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aivodji, Aisha Alaagib, Sébastien Gambis, and Nicolas Papernot. 2022. Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems* 35 (2022), 14170–14182.
- [171] Manali Sharma and Mustafa Bilgic. 2017. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery* 31 (2017), 164–202.
- [172] Rahul Sharma, Sergey Redyuk, Sumantrak Mukherjee, Andrea Sipka, Sebastian Vollmer, and David Selby. 2024. X Hacking: The Threat of Misguided AutoML. *arXiv preprint arXiv:2401.08513* (2024).
- [173] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.
- [174] Jan Simson, Florian Pfisterer, and Christoph Kern. 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.
- [175] D Singh. 1993. Godel’s Incompleteness theorem in particular reference to Artificial Intelligence. In *Godel’s Theorems-Proceedings Of The First International Symposium*. World Scientific, 219.
- [176] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. 2017. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE transactions on evolutionary computation* 22, 2 (2017), 276–295.
- [177] Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. [n. d.]. Rethinking Aleatoric and Epistemic Uncertainty. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.
- [178] Gavin Smith, Roberto Mansilla, and James Gouling. 2020. Model class reliance for random forests. *Advances in Neural Information Processing Systems* 33 (2020), 22305–22315.
- [179] Kacper Sokol, Meelis Kull, Jeffrey Chan, and Flora Salim. 2024. Cross-model Fairness: Empirical Study of Fairness and Ethics Under Model Multiplicity. *ACM Journal on Responsible Computing* 1, 3 (2024), 1–27.
- [180] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. 2022. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13699–13708.
- [181] Cătălin Soreanu and Lavinia German. 2022. 5. Visual Communication in Cultural Media. The Rashomon Effect in the Image Globalization Paradigm. *Review of Artistic Education* 26, 1 (2022), 177–185.
- [182] Sarath Sreedharan, Subbarao Kambhampati, et al. 2018. Handling model uncertainty and multiplicity in explanations via model reconciliation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 28. 518–526.
- [183] Yiyang Sun, Tong Wang, and Cynthia Rudin. [n. d.]. Improving Decision Sparsity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [184] Hanxiao Tan. 2024. Evaluating Explanation Robustness to Model Pruning. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
- [185] Damien Teney, Maxime Peyrard, and Ehsan Abbasnejad. 2022. Predicting is not understanding: Recognizing and addressing underspecification in machine learning. In *European Conference on Computer Vision*. Springer, 458–476.
- [186] Christopher Tindale. 2016. Collateral beliefs and the Rashomon effect. *Cogency* 8, 2 (2016), 125–145.
- [187] Tom Trabasso. 2018. Whose memory is it? The social context of remembering. In *Memory for everyday and emotional events*. Routledge, 429–443.
- [188] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems* 33 (2020), 1633–1645.
- [189] Dustin Tran, Jasper Snoek, and Balaji Lakshminarayanan. 2020. Practical uncertainty estimation and out-of-distribution robustness in deep learning. *NeurIPS Tutorial, Google Brain* (2020).
- [190] Theja Tulabandhula and Cynthia Rudin. 2014. Robust Optimization using Machine Learning for Uncertainty Sets. *ECML/PKDD 2014* (2014), 121.
- [191] Masao Ueki and Yoshinori Kawasaki. 2013. Multiple choice from competing regression models under multicollinearity based on standardized update. *Computational Statistics & Data Analysis* 63 (2013), 31–41.
- [192] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*. PMLR, 3459–3467.
- [193] Thomas Veran, Pierre-Edouard Portier, and François Fouquet. 2023. Interpretable hierarchical symbolic regression for safety-critical systems with an application to highway crash prediction. *Engineering Applications of Artificial Intelligence* 117 (2023), 105534.
- [194] Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766* (2024).
- [195] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–45.
- [196] Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning*. PMLR, 10882–10893.
- [197] Jue Wang and Qing Tao. 2008. Machine learning: The state of the art. *IEEE Intelligent Systems* 23, 6 (2008), 49–55.
- [198] Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. 2022. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 875–884.

- [199] Jamelle Watson-Daniels. 2024. *The Roads Not Taken: Model Multiplicity in Machine Learning*. Ph. D. Dissertation. Harvard University Graduate School of Arts and Sciences.
- [200] Jamelle Watson-Daniels, Solon Barocas, Jake M Hofman, and Alexandra Chouldechova. 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.
- [201] Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D’Amour, Carol Long, David C Parkes, and Berk Ustun. 2024. Predictive churn with the set of good models. *arXiv preprint arXiv:2402.07745* (2024).
- [202] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10306–10314.
- [203] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. 2021. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems* 34 (2021), 7054–7067.
- [204] Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209* (2024).
- [205] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems* 35 (2022), 14071–14084.
- [206] Yazhou Yang and Marco Loog. 2016. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2646–2651.
- [207] Peiliang Zhang, Huan Wang, Nikhil Naik, Caiming Xiong, et al. 2020. DIME: An information-theoretic difficulty measure for AI datasets. In *NeurIPS 2020 Workshop: Deep Learning through Information Geometry*.
- [208] Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. 2024. An Introduction to Bilevel Optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine* 41, 1 (2024), 38–59.
- [209] Chudi Zhong. 2024. *Interpretability and Multiplicity: a Path to Trustworthy Machine Learning*. Ph. D. Dissertation. Duke University.
- [210] Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. 2024. Exploring and interacting with the set of good sparse generalized additive models. *Advances in neural information processing systems* 36 (2024).
- [211] Gianluca Zuin, Felipe Marcelino, Lucas Borges, João Couto, Victor Jorge, Mychell Laurindo, Glaucio Barcelos, Marcio Cunha, Valdeci Alvarenga, Henrique Rodrigues, et al. 2021. Predicting Heating Sliver in Duplex Stainless Steels Manufacturing through Rashomon Sets. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

## A Systematic Literature Review

We conducted a systematic literature review to collect and analyze works related to multiplicity. Our process involved the following steps:

- (1) **Paper Collection:** We indexed various online repositories using a set of predefined search terms. Details of the search are provided in §A.1. After filtering for duplicates across various search terms and repositories, we retained a total of 339 papers.
- (2) **Manual Filtering:** We manually reviewed the remaining papers to exclude those not relevant to our study. The exact inclusion criteria are noted in §A.2. This manual filtering reduced the dataset to 80 papers.
- (3) **Manual Tagging:** Finally, each paper was assigned tags to identify and emphasize the specific aspects of multiplicity it addressed. The set of tags used and categorization rules are discussed in §A.3.

After filtering and tagging the papers, we analyzed various statistics, as shown in the main text. The following sections provide a detailed account of each step of the systematic review process.

### A.1 Paper Collection

We indexed and collected papers using the search terms ‘rashomon’, ‘model multiplicity’, and ‘set of good models’ from two primary sources: DBLP<sup>1</sup> and the ACM Digital Library<sup>2</sup>. To ensure comprehensive coverage, we used the extended search in ‘The ACM Guide to Computing Literature’ within the ACM Digital Library. The end date for our search was 31 Dec 2024, and no restrictions were put on the starting date. Together, DBLP and the ACM Digital Library provide coverage of major machine learning conferences (e.g., NeurIPS, ICML, ICLR, AAAI, ACL, NAACL), leading venues

<sup>1</sup><https://dblp.org/>

<sup>2</sup><https://dl.acm.org/>



Search Term	Number of Papers		
	DBLP	ACM Digital Library	Combined (Duplicates Removed)
'rashomon'	34	156	178
'model multiplicity'	66	95	153
'set of good models'	11	28	38
Total Number of Papers (Duplicates Removed):			339
<b>Total Number of Papers (After Manual Filtering):</b>			<b>80</b>

Table 3. Paper collection statistics.

Inclusion Criteria	Exclusion Criteria
<ul style="list-style-type: none"> <li>- Papers whose central contribution was deeply intertwined with the Rashomon effect or multiplicity.</li> <li>- Position papers and surveys that devoted a considerable amount of space to the Rashomon effect or multiplicity (We used one complete section worth of space as our threshold).</li> </ul>	<ul style="list-style-type: none"> <li>- Papers that mentioned the Rashomon effect, but did not engage with it in their main contributions.</li> <li>- Papers exploring the Rashomon effect or model multiplicity in domains outside machine learning. The term 'model' in these cases did not refer to machine learning models.</li> <li>- Commentaries, extended abstracts, tutorials, complete conference proceedings, and other indexed documents that are not considered research papers.</li> <li>- Duplicates for papers which exist in various repositories under different names, and hence could not be filtered automatically.</li> </ul>

Table 4. Inclusion and exclusion criteria for manual filtering.

focused on responsible AI (e.g., FAccT, AIES, EAAMO), and archival repositories such as arXiv. The exact number of papers collected from each source and search query is detailed in Table 3.

## A.2 Manual Filtering

After removing duplicates, we had 339 papers in our dataset. Each paper was then manually reviewed to determine its relevance to our survey. Our focus was on works that deeply engaged with the Rashomon effect and multiplicity within the context of machine learning. Thus, papers examining the Rashomon effect outside of machine learning, investigating multiplicity in other domains, or briefly mentioning the Rashomon effect without further exploring it, were excluded. The detailed inclusion and exclusion criteria are outlined in Table 4.

## A.3 Manual Tagging

Finally, after manual filtering, we were left with 80 papers. Each paper was then tagged based on their contributions in the context of multiplicity in machine learning. The tags were created by the authors after a preliminary review of the papers during the manual filtering step. Strict rules were defined for each tag, as shown in Table 5. Every paper was assigned all applicable tags through this manual tagging process. These 80 tagged papers formed the basis of the statistics presented in Fig 1 of the main text.

<b>Tag</b>	<b>Rule</b>
Rashomon Effect in ML	Papers that provide fundamental insights into the Rashomon effect in ML, including its causes, forms of manifestation, broader characteristics, etc.
Rashomon Set Exploration	Papers that focus on the problem of enumerating the Rashomon set, including papers which highlight properties of the Rashomon set. This does not include papers that already assume access to the Rashomon set.
Better Models and Ensembles	Papers that use the Rashomon set to find better models, or combine multiple models from the Rashomon set to improve certain objectives. Thus, any work that takes advantage of the flexibility the Rashomon set provides.
Hacking Metrics	Papers that highlight the negatives of underspecification, using the Rashomon effect to hack existing metrics, checks, or regulations.
Responsible AI	Papers that explicitly engage with the harms or benefits of the Rashomon effect in the context of Responsible AI. Papers whose central contribution is related to Responsible AI, but do not engage with that aspect of their impact, were not included here. For example, papers tackling explanation multiplicity strictly from a technical point of view were not given this tag.
Application	Papers whose central contribution is limited to a particular application.
Survey	Position papers and surveys that either provide an overarching discussion on multiplicity or place multiplicity in a broader context.
Predictive Multiplicity	Papers that focus on predictive multiplicity, i.e., when models in the Rashomon set have varying predictions.
Explanation Multiplicity	Papers that focus on explanation multiplicity, i.e., when models in the Rashomon set have varying explanations.
Other Multiplicity	Papers that focus on any other form of multiplicity beyond predictive or explanation multiplicity. This can include fairness multiplicity, OOD robustness multiplicity, model complexity multiplicity, feature interaction multiplicity, etc.

Table 5. Rules for each tag in the manual tagging step.