We need to explain our explanations

Venetia Pliatsika
Computer Science and Engineering
New York University
New York, NY
venetia@nyu.edu

Abstract

Shapley values, derived from cooperative game theory, are widely used in machine learning to design, audit, and recommend recourse for complex models. A wide variety of implementations has resulted in explanation multiplicity, the phenomenon where multiple contradictory Shapley value explanations can be derived for the same model and data instance. In this position paper, we argue that an explanation is mathematically underspecified without a formal definition of context of use and a corresponding disclosure of technical assumptions. To enable both goals, we introduce a generalized definition of Shapley values and use it to reframe multiplicity as a parametrization problem. Using this unified parametrization, we propose a roadmap for the entire explanation lifecycle, calling focus on two primary research directions. First, we posit that *context of use is a formal input to the explanatory process*. As such, we require rigorous guidelines to map stakeholder questions to specific mathematical parameters. Second, we argue that an explanation is incomplete without a *meta-explanation*: a structured disclosure of the underlying assumptions and implementation choices used in its construction. As a step towards this second direction, we propose the Shapley Value Explainability Card. Finally, we demonstrate through several examples how each parameter choice implies specific mathematical assumptions about the data or the explanation's intended use, and provide a detailed call to action towards a more transparent and context-aware explanation lifecycle.

1 Introduction

Shapley values [40] have emerged as a dominant framework for feature attribution in machine learning, largely due to their axiomatic foundation in cooperative game theory. However, their theoretical promise of uniqueness is increasingly contradicted by practical ambiguity, which can result in misuse [11]. A growing body of literature demonstrates that applying different Shapley value implementations to the same model and unit can yield significantly differing—and often contradictory—explanations [45, 32, 48, 9, 5, 23, 11, 42, 26, 10]. This phenomenon of "explanation multiplicity" [45] (or "Rashomon effect" [7, 42]) has sparked a contentious debate regarding trustworthiness: if a single unit can be explained in multiple ways, which explanation is correct?

Recent research suggests that these discrepancies are often attributable to specific *modeling choices* made during the explanation process [45, 23, 9, 29, 32]. For example, correlated features that are not used by the model can have significant or zero importance depending on the sampling methodology [9, 29]. This demonstrates that feature importance is not an intrinsic property of the model-unit pair, but rather a normative outcome of the underlying choices for the explanation.

In this work, we first express these modeling choices as parametrization of the Shapley game, and frame explanation multiplicity as the result of different parameter choices. Then we use this parametrization to propose a new explanation lifecycle, Fig. 1. We argue that when different

Preprint.

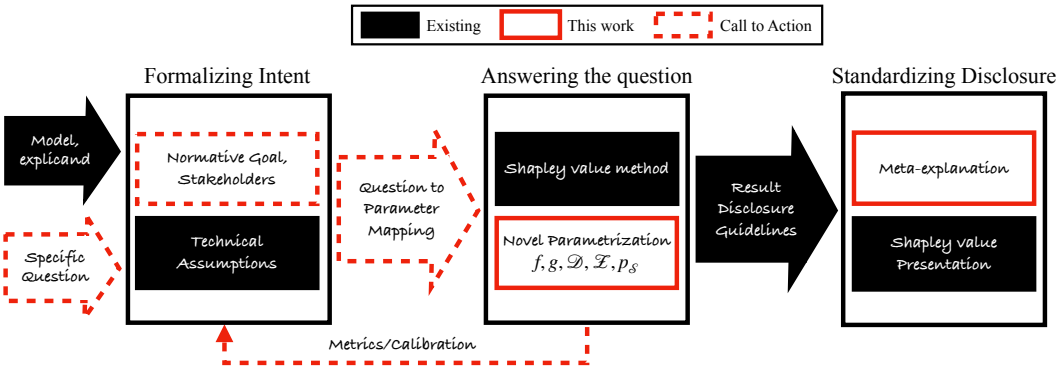
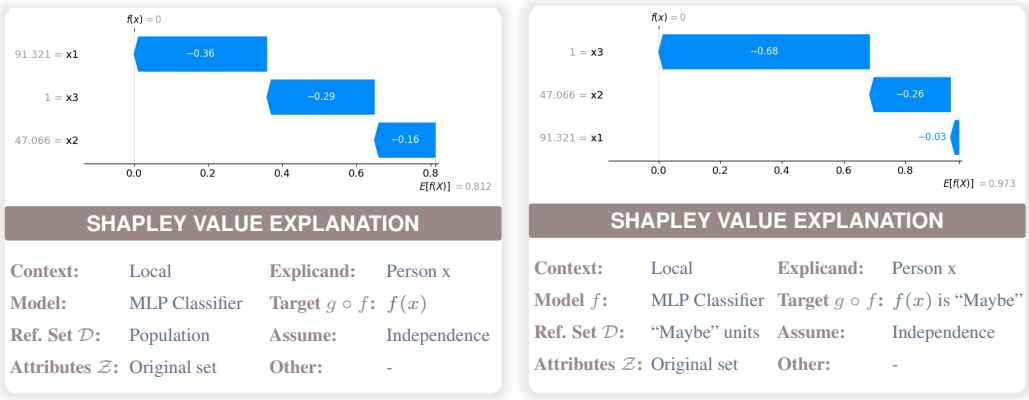


Figure 1: The proposed explanation pipeline that uses the unified definition to define the Shapley parameters and the Shapley explainability cards (meta-explanation). We issue a call to action for many parts of the pipeline, most notably mapping the explainability context to the mathematical parameters.



(a) Explaining unit x using the entire dataset and the model

(b) Explaining unit x using a subset of the dataset and a composite profit function

Figure 2: Different Shapley value games explain different questions for the same model and person.

parametrization can lead to several significantly different but valid explanations, the community must abandon the search for a singular unique Shapley value in favor of documented normative alignment. Specifically, **we propose that we need to 1) formally describe the relationship between parameter choice and context of use of the explanation, and 2) disclose the explanation design choices to the explanation recipient.**

We make the following contributions: 1) We introduce a unified definition of Shapley values that integrates existing disjoint methods into a single parametric framework, 2) We demonstrate, through various game formulations, that feature importance is relative to the choice of parameters $(f, g, \mathcal{D}, \mathcal{Z}, p_{\mathcal{S}})$ and that different valid choices correspond to semantically distinct questions, 3) We propose a new explanation lifecycle (Fig. 1) and outline a call to action for future research focused on mapping these parameters to specific stakeholder goals and normative contexts, and 4) we introduce the concept of meta-explanations and, as an initial solution to the problem of disclosure, we propose the Shapley value Explainability Card (Fig. 2), a documentation standard inspired by Model Cards [33], to communicate these parameters to stakeholders.

2 A unified definition for Shapley values

Existing Shapley value implementations are often presented as distinct algorithms, yet they can be unified as specific configurations of a single mathematical definition. We propose a generalized definition that extends existing work [6, 32]. See also Algorithm 1.

Algorithm 1 Calculating Shapley values

Require: Model f, profit function g, hybrid data distributions $p_{\mathcal{S}}^{\mathcal{D},X_{obs}}$ for all $\mathcal{S}$ where $\mathcal{S} \subseteq \mathcal{Z}$, parameters $\mathcal{A}$ and m^{-1}.

$\phi(\mathbf{x}) = \langle 0, \ldots, 0 \rangle$
for $i \in \mathcal{Z}$ **do**
  for $\mathcal{S} \subseteq \mathcal{Z} \setminus \{i\}$ **do**
    $\mathbf{U}_1 \sim p_{\mathcal{S} \cup i}$
    $\mathbf{U}_2 \sim p_{\mathcal{S}}$
    $\mathbf{X}_1 = m^{-1}(\mathbf{U}_1)$
    $\mathbf{X}_2 = m^{-1}(\mathbf{U}_2)$
    $\phi_{i_{\mathcal{S}}}(\mathbf{v}) = g(f, \mathbf{U}_1, \mathcal{A}) - g(f, \mathbf{U}_2, \mathcal{A})$
    $\phi_i(\mathbf{v}) = \phi_i(\mathbf{v}) + \frac{1}{|\mathcal{Z}|} \frac{1}{\binom{|\mathcal{Z}|-1}{|S|}} \phi_{i_{\mathcal{S}}}(\mathbf{v})$
  end for
end for
Return $\phi(\mathbf{v})$

Definition 2.1 (Shapley value)**.** Given a model f, some profit function $g \circ f$, a background dataset $\mathcal{D}$ consisting of features $\mathcal{X}$, a set of features to be explained $\mathcal{Z}$, and an invertible function $m : \mathcal{Z} \to \mathcal{X}$ that maps the elements of $\mathcal{Z}$ to elements of $\mathcal{X}$. For an observed unit x_{obs}, we define distributions $p_{\mathcal{S}}^{\mathcal{D},x_{obs}}$ for each $\mathcal{S} \subseteq \mathcal{Z}$, denoted as $p_{\mathcal{S}}$ for brevity when $\mathcal{D}$ and x_{obs} are clear from context. Then the contribution of each feature $\mathcal{Z}$ to the outcome value $g(f(x_{obs}))$ of is the following, where $\mathcal{A}$ is a set of parameters, and $n = |Z|$.

$$\phi_i = \sum_{\mathcal{S} \subseteq \mathcal{Z} \setminus i} \frac{|\mathcal{S}|!(n - |\mathcal{S}| - 1)!}{n!} \left(\mathop{\mathbb{E}}_{x \sim p_{\mathcal{S} \cup i}} [g(f, x, \mathcal{A})] - \mathop{\mathbb{E}}_{x \sim p_{\mathcal{S}}} [g(f, x, \mathcal{A})] \right) \tag{1}$$

The unified Shapley value definition allows us to further define a specific Shapley game and a specific Shapley question.

Definition 2.2 (The Shapley Game and the Shapley Question)**.** Every instantiation of $\{f, g, \mathcal{D}, \mathcal{Z}, \mathcal{X}, p_{\mathcal{S}}\}$ defines a specific *Shapley Game*, which answers a specific *Shapley Question*: "How does the presence of attributes $\mathcal{Z}$ contribute to the unit x_{obs} achieving the outcome $g(f(x_{obs}))$, relative to the expected outcome of hypothetical reference units drawn $p_{\mathcal{S}}^{\mathcal{D},x_{obs}}$?"

Using the generalized definition, we situate existing literature within a unified parameter space. We highlight how disparate modeling choices give rise to the phenomenon of explanation multiplicity.

2.1 Fill-in methods

Shapley values are computed over coalitions (subsets) of varying sizes. While one could train a unique model for every possible coalition [44], we typically aim to explain a single model defined over inputs of a fixed length. We therefore must "fill-in" missing values [11], a process that requires assumptions about the characteristics of hypothetical units—namely, which values features can take and how they interact. Following [32], we frame disparate fill-in methods as specific ways to parametrize the background data $\mathcal{D}$ and the hybrid data distributions $p_{\mathcal{S}}^{\mathcal{D},x_{obs}}$.

2.1.1 Background data

The background data $\mathcal{D}$, typically represented as an empirical dataset where $x_{obs} \in \mathcal{D}$, or as a Data Generating Process (DGP) for causal frameworks, constrains $p_{\mathcal{S}}^{\mathcal{D},x_{obs}}$. Specifically, $\mathcal{D}$ acts as a functional constraint that dictates the support and range of perturbations available to the explainer, regardless of the specific distributional assumptions eventually adopted.

While subselecting background data is often employed to facilitate tractable approximations of the Shapley value, the choice to restrict $\mathcal{D}$ frequently serves a normative purpose, to curate a specific contrastive signal [32, 48]. For example, explaining why a student received a 'B+', using as background data students who received an 'A-' [32], or evaluating the 15-year mortality risk

of a 75-year-old male by subselecting a cohort of age- and gender-matched peers instead of the general-population, a baseline which is clinically irrelevant [10]. So, because these choices are dictated by context-specific logic rather than mathematical necessity, they represent a hidden "degree of freedom" that must be explicitly documented.

The following lemma characterizes how background data determines the total distance between explanations parameterized by distinct datasets.

Lemma 2.3. *Let f be a model, g a profit function, and $\phi(f, x, \mathcal{D}_1)$ and $\phi(f, x, \mathcal{D}_2)$ the Shapley value explanations for observation x produced using background data $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Their aggregate difference in feature attributions is given by:*

$$d_{D_1,D_2}(x) = \sum_i \phi_i(f, x, \mathcal{D}_1) - \sum_i \phi_i(f, x, \mathcal{D}_2) = \mathbb{E}_x^{\mathcal{D}_2}[f(x)] - \mathbb{E}_x^{\mathcal{D}_1}[f(x)] \tag{2}$$

Proof. Proof follows from the efficiency property [34, 45]: $\sum_i \phi_i(f, x, \mathcal{D}) = f(x) - \mathbb{E}_x f(x)$. □

Impact on explanation. We see that the aggregate difference in feature contributions between the two explanations derived from distinct background subsets is equivalent to the difference in the means of those subsets. While this difference may be negligible under unbiased sampling, it becomes a primary driver of explanation multiplicity when subsets are selected purposefully to satisfy a contrastive property. In the grading example, comparing a student against an 'F' cohort versus an 'A+' cohort creates a fundamental shift in the reference mean. Such a selection would invert the signs of the feature contributions and potentially their magnitudes. Consequently, the "important" features might be fundamentally different between the two explanations based solely on the choice of $\mathcal{D}$.

2.1.2 Hybrid data distributions

Constructing hybrid units–mixtures of observed and reference features–requires defining a distribution to fill-in the complement set $\bar{\mathcal{S}}$. Existing literature primarily considers five classes of distributions: uniform [43], marginal [12, 30, 14, 31], product of marginals [16], conditional [2, 21, 36], and—where a causal graph is known—interventional [23].We extend the framework from [32] to incorporate the interventional distribution.

Formally, given a unit x_{obs}, background data $\mathcal{D}$, and a subset of features $\mathcal{S}$, we must define values for the complement set $\bar{\mathcal{S}} = \mathcal{X} \setminus \mathcal{S}$. To estimate the expected value $\mathbb{E}_{x \sim p_{\mathcal{S}}^{\mathcal{D}, x_{obs}}}[f(x)]$, we generate hybrid units where features in $\mathcal{S}$ are set to $x_{obs,\mathcal{S}}$ (or $do(\mathcal{X}_{\mathcal{S}} = x_{obs,\mathcal{S}})$). The features in $\bar{\mathcal{S}}$ are then sampled from a distribution $p_{\bar{\mathcal{S}}}^{\mathcal{D}, x_{obs}}$ derived from one of the aforementioned strategies.

A frequent simplification uses a single "baseline unit" to populate $\bar{\mathcal{S}}$, effectively replacing $\mathcal{D}$ with a single unit; this represents a special case of the marginal distribution. Common "neutral" choices include the all-zeroes vector or the mean feature value [30, 19]. More recently, "dynamic" baselines have been introduced to facilitate pairwise comparisons between specific instances [39, 32].

Impact on explanation. The diverse assumptions of these distributions have sparked intense debate [28, 45]. Conditional distributions capture data correlations, attributing importance even to features unused by the model, whereas marginal distributions possibly attributes model-specific importance [32, 29, 9]. Uniform, marginal, and product-of-marginal approaches differ in the weight assigned to outliers during hybrid unit construction [32]. Only the interventional distribution respects causality, providing both direct and indirect effects [23], though some argue the marginal is closer to the interventional than the conditional [26, 29]. These differences make distributional assumptions pivotal for the resulting explanation.

2.2 Explanation Features

Existing methodologies often advocate for an attribution space $\mathcal{Z}$ distinct from the model's input space $\mathcal{X}$. We formalize this via a mapping $m : \mathcal{Z} \to \mathcal{X}$, where explanations are calculated over $\mathcal{Z}$ but model evaluations occur in $\mathcal{X}$.

[29] utilizes such mappings to group correlated features while using marginal distributions. For example, systolic and diastolic pressures may be aggregated into a single "blood pressure" feature

to reflect that medical interventions typically affect these variables simultaneously. In our notation, $z_i \in \mathcal{Z}$ maps to the tuple $(x_j, \ldots, x_k) \subseteq \mathcal{X}$, ensuring that hybrid instances remain in the space of $\mathcal{X}$ while providing semantically coherent concept-level explanations that are always sampled jointly.

We suggest a similar imperative for one-hot encoded features. Mapping categorical variables in $\mathcal{Z}$ to their encoded representations in $\mathcal{X}$ preserves interpretability, prevents unnecessary feature proliferation, and ensures distributional consistency by avoiding nonsensical hybrid states where multiple mutually exclusive categories are simultaneously active.

Finally, in recourse, attributing importance to non-actionable attributes like age may be undesirable [47]. Using the mapping to make these variables invariant, we restrict attribution to actionable dimensions.

Impact on explanation. These choices—grouping, re-encoding, and exclusion-fundamentally alter the list of features that are explained, and therefore the magnitude of the new features.

2.3 Profit functions

Recent work argues that many explainability questions cannot be addressed by explaining model outputs $f(x)$ alone, proposing instead to explain a composite profit function $g \circ f$ [6, 16, 39, 13, 24]. While g commonly defaults to the identity function, it can represent specific latent consequences or other functions defined over the model output.

For example, in ranking, profit functions, based on the rank of a unit defined over the score of all units, are used to explain ranked outputs [39, 13, 24]. For these methods, we could set g to be the function of a unit's rank, setting $\mathcal{A}$ to be the score $f(x)$, and rank of all other observed units in the dataset. Further, [16, 6] introduce profit functions for a variety of questions, such as intergroup differences. Our framework accommodates these by allowing g to handle sets of observations $x_i \subseteq X$, achieved by setting $p_{\mathcal{S}}^{\mathcal{D}}$ to include hybrid units for all x_i, and ensuring that empirical averaging occurs within the profit function to accommodate different averaging strategies per profit function (see Algorithm 1 and the specific profit functions of these works).

Returning to the grade example from Section 2.1.1, consider a student who received a B+. By explaining the function $g(f(x)) = \mathbb{1}_{f(x)=A-}(x)$ instead of $f(x)$, we shift the inquiry from the label changing to any grade, to the grade being within the decision boundary of A-; contributions to receiving any other grade are rendered irrelevant, thus isolating the features that can induce A-. Crucially, this is distinct from selecting a reference set $\mathcal{D}$ consisting of students who received an A-. The reference set determines the source of possible feature values used to construct hybrid units, providing values that, in some combination, were graded as A-. The profit function, conversely, determines the evaluation criterion—dictating whether those hybrid units count as a "success."

Impact on explanation. Altering the profit function does not merely shift the mathematical baseline; it fundamentally changes the unit, magnitude, and sign of the attribution. Normatively, because the choice of g defines the unit of the question being asked, it must be explicitly disclosed for the resulting explanation to be valid and comprehensible.

2.4 Higher-order Shapley values

Our framework naturally incorporates Shapley value interactions (SVI) [46] or any measure based on a weighted sum of differences, such as the Banzhaf Index [4, 16]. We introduce a generalized definition that flexibly accommodates various indices by modifying the target set $\mathcal{T}$, weights $w_{\mathcal{S}}$, and difference function $\Delta_{\mathcal{S}}$, (remember $n = |\mathcal{Z}|$):

$$\phi_{\mathcal{T}} = \sum_{\mathcal{S} \subseteq \mathcal{Z} \setminus \mathcal{T}} w_{\mathcal{S}} \Delta_{\mathcal{S}} \tag{3}$$

Shapley: $\mathcal{T} = \{i\}$, $w_{\mathcal{S}} = \frac{|\mathcal{S}|!(n-|\mathcal{S}|-1)!}{n!}$, and $\Delta_{\mathcal{S}} = \mathbb{E}_{x \sim p_{\mathcal{S} \cup \{i\}}}[g(f, x, \mathcal{A})] - \mathbb{E}_{x \sim p_{\mathcal{S}}}[g(f, x, \mathcal{A})]$.

Banzhaf Index: $\mathcal{T} = \{i\}$, $w_{\mathcal{S}} = \frac{1}{2^{n-1}}$, and $\Delta_{\mathcal{S}}$ matches Shapley.

SVI: $\mathcal{T} \subseteq \mathcal{Z}$, $w_{\mathcal{S}} = \frac{(n-|\mathcal{T}|-|\mathcal{S}|)!|\mathcal{S}|!}{(n-|\mathcal{T}|+1)!}$, and $\Delta_{\mathcal{S}} = \sum_{\mathcal{W} \subseteq \mathcal{T}} (-1)^{|\mathcal{W}|-|\mathcal{T}|} \mathbb{E}_{x \sim p_{\mathcal{S} \cup \mathcal{W}}}[g(f, x, \mathcal{A})]$.

Impact on Explanation. While we do not include $\mathcal{T}$, $w_{\mathcal{S}}$, and $\Delta_{\mathcal{S}}$ in definition 2.1 to minimize the notation and keep the position simple, the method being used has a large impact on the explanations. For example, SVI are able to account for feature interactions of up to size $|\mathcal{T}|$.

2.5 Approximations

Analogous to higher-order indices, Shapley value approximation methods (for example [30, 10, 8, 35, 27, 31, 3, 45, 1]) function as aggregation procedures that take the parameters of Definition 2.1 as input to estimate feature contributions. These methods introduce various forms of approximation error and new parameters depending on their implementation—such as the number of coalition samples m in Monte Carlo approaches [8] or the specific distributional assumptions used to approximate conditional expectations [9, 2]. While these auxiliary parameters are omitted from the core definition for brevity, approximations are critical components of the explanatory metadata and are thus included in our proposed Shapley Card.

Impact on Explanation. Approximation methods introduce inherent uncertainty into the resulting attribution. Consequently, the choice of algorithm and, more importantly, the expected error bounds must be disclosed to the recipient to ensure the explanation's reliability and prevent over-interpretation of statistically insignificant results.

2.6 Complex games

Shapley values extend beyond standard model explanations to address more complex inquiries, such as identifying biased outcomes using pattern mining [37] or explaining pairwise preference learning [25]. Our framework naturally accommodates these complex formulations. We provide more details on how our framework can represent these two specific examples below.

In [25], they examine problems that do not adhere to the rankability assumption, specifically, scenarios where no total order exists over the items, but rather only pairwise preferences. To explain a preference relation between two units x^l and x^r, they construct a composite unit (x^l, x^r) and perturb the features of both items simultaneously—for a set $\mathcal{S}$, both $x^l_{\bar{\mathcal{S}}}$ and $x^r_{\bar{\mathcal{S}}}$ are altered at the same time. We can model this through the attribute space $\mathcal{Z}$ using the feature grouping mechanism from [29]. Specifically, we define the explainable attributes as tuples $\mathcal{Z}_i = (\mathcal{X}^l_i, \mathcal{X}^r_i)$, forcing the simultaneous perturbation required.

In [37], they employ pattern mining to identify biased subgroups defined by specific feature-value pairs (e.g., $\{\text{age} > 25, \text{sex} = \text{Male}\}$). After identifying itemsets with anomalous outcomes, they use Shapley values to determine which term in the itemset definition drives the disparity. In this setting, the "absence" of a feature corresponds to removing a constraint from the itemset definition. Consequently, no sampling is required. We model this deterministically by defining $p_{\mathcal{S}}$ such that it yields the modified itemset (the subset of constraints) with probability 1.

Impact on Explanation. We argue that as the complexity of the game increases, explicit documentation of the explanatory parameters becomes indispensable.

3 The explanation lifecycle: a call to action

The primary goal of this work is a paradigm shift in the explanation lifecycle, Fig. 1. We argue that context of use must be a formal input to the explanatory process, driving both the mathematical production of the attribution and its subsequent disclosure. Context can include diverse objectives such as model development, auditing, recourse recommendation, and individual decision-level explanations. By explicitly incorporating context, we disambiguate *explanation multiplicity*–the lack of a unique explanation per model and unit. This disclosure increases transparency and clarifies the transportability of the explanation across different domains.

3.1 Formalizing Intent

To provide a meaningful explanation, practitioners must first *map the normative context to specific mathematical parameters*. We call this process defining the "Shapley question." Current work focuses heavily on the computational properties of estimation techniques. We urge the community to expand

Table 1: Components of the Shapley Card.

Parameter	Notation	Technical Specification	Meaning
Intended Context		Normative definition of context	Why was this explanation created?
Explicand	x_{obs}	Unit(s) being explained	Who is this explanation about?
Model	$f, \mathcal{X}$	Model to be interpreted and input space	How are the outcomes produced?
Target	$g \circ f$	Profit function (Units/Scale)	What specific outcome is being explained?
Reference	$\mathcal{D}$	Background data	Who are we comparing against?
Attributes	$\mathcal{Z}$	Feature grouping/mapping	What are the players in the coalition?
Assumptions	$p_{\mathcal{S}}$	Distributional assumptions	How do we generate hybrid units?
Other		Implementation / Approximation	Is the value estimated? (what is the error)?

this rigor to the normative implications of all parameters. Future work must establish a formal taxonomy of standard configurations tailored to specific stakeholder goals. For example, auditing may require the population as the reference set $\mathcal{D}$ to detect systemic bias, whereas individual recourse may require a reference set of"similar neighbors" to suggest attainable changes. Furthermore, research is required to identify the bounds of manipulation; we must determine which parameter combinations are robust for a given context and which are brittle or prone to adversarial exploitation [41].

3.2 Answering the question

Once the parameters are set, the resulting Shapley explanations can be evaluated. Existing evaluation metrics typically measure "faithfulness" to the model [5, 15]. However, if multiple faithful explanations exist, we need new metrics that measure *alignment*: does the chosen configuration actually answer the specific question the user intended to ask? How do the different explanations compare?

3.3 Standardizing Disclosure: the Shapley Card

Because distinct parameter configurations could correspond to equally valid yet semantically distinct questions, presenting a Shapley value in isolation is not merely incomplete—it is potentially misleading. To address this ambiguity, we propose the **Shapley Card**—a structured meta-explanation documentation framework inspired by Model Cards [33] and Datasheets for Datasets [22]. By explicitly reporting the parameter choices that define the underlying cooperative game (Table 1) —The Shapley Card provides the necessary context to transform a feature importance score into a transparent, interpretable artifact. However, the Card is only a starting point. We call on the community to design meta-explanations that go beyond raw parameters, creating interfaces appropriate for stakeholders with varying technical literacy. Figure 2 shows an example of a Shapley Card.

4 Case Studies

This section presents case studies across synthetic and real-world datasets, spanning both tabular and image data. The examples serve a dual purpose: first, we demonstrate the practical impact of explanation multiplicity, a well-documented challenge in the literature [45, 32, 48, 9, 5, 23, 11, 42, 26, 10], second, we provide an initial paradigm for mapping explanation context to specific Shapley Questions. We provide a simple case study using Partial Dependence Plots (PDPs) [20] in Appendix D.

4.1 Case Study 1: Synthetic Dataset - Average vs. Targeted Local Explanation

Dataset - Survival show auditions We created an artificial dataset of 28,000 datapoints inspired by [48]. The task is to get an audition for a survival show. The dataset consists of three features X_1, X_2, and X_3, which can be thought of as "muscle mass", "height", and "gender". We have three classification outcomes, "no", "maybe", and "yes", indicating whether someone will be auditioned or not. We use different features for deciding the outcome per gender. We create the dataset as follows.

We define X_1 as a Gaussian mixture $X_1 \sim 0.29\mathcal{N}(25, 5^2) + 0.42\mathcal{N}(75, 5^2) + 0.29\mathcal{N}(125, 5^2)$. The variable $X_3 \sim \text{Bernoulli}(0.5)$ serves as a gender indicator where 0 represents men and 1 represents women. Finally, X_2 is a conditional Gaussian mixture dependent on X_3: for women ($X_3 = 1$),

$X_2 \sim 0.29\mathcal{N}(40, 5^2) + 0.42\mathcal{N}(45, 5^2) + 0.29\mathcal{N}(55, 5^2)$, whereas for men ($X_3 = 0$), the distribution shifts to $X_2 \sim 0.29\mathcal{N}(40, 5^2) + 0.42\mathcal{N}(50, 5^2) + 0.29\mathcal{N}(60, 5^2)$.

Whether a unit will audition is determined based on the following rules, and illustrated in Figure 7 in the Appendix. We can quickly note that height (X_2) matters only for women, and that there are women who would get a better outcome if they were men.

$$no : X_1 \leq 50 \tag{4}$$
$$no : 50 < X_1 \leq 100 \text{ and } X_3 = 1 \text{ and } X_2 \leq 50 \tag{5}$$
$$maybe : 50 < X_1 \leq 100 \text{ and } X_3 = 0 \tag{6}$$
$$maybe : 50 < X_1 \leq 100 \text{ and } X_3 = 1 \text{ and } X_2 > 50 \tag{7}$$
$$yes : X_1 > 100 \tag{8}$$

Model We train-test split 70-30% of the data, and train a Multi-Layer Perceptron Classifier, with a maximum of 300 iterations. The accuracy is 0.993, and the confusion matrix is shown below in Figure 3a.

Context. Let's suppose we are the agent of contestant x, shown in Figure 3b. The contestant is a woman ($X_3 = 1$) with $X_1 = 91$ and $X_2 = 47$. The contestant received the "no" outcome as dictated by rule 5. We are interested in finding out why our contestant wasn't wait-listed.

Parameter Setting 1. In Figure 2a, we see the explanation we would get out-of-the-box using SHAP on this dataset [1]. This simply means that we implicitly chose values for all variables in Definition 2.1. Specifically, for this dataset, we chose $g = f$ and $\mathcal{Z} = \mathcal{X}$, and $p_S^{\mathcal{D}}$ are created using the marginal fill-in method, and $\mathcal{D}$ is the entire training set. According to Definition 2.2 then, Figure 2a corresponds to "How would the features $\{X_1, X_2, X_3\}$ contribute to unit x_a receiving outcome *no*, relative to the average outcome of other units drawn assuming marginally independent features with probability distributions defined using the values of the entire training population." While this question and the corresponding explanation are perfectly valid, we will examine another option.

Parameter Setting 2. The agent of this unit wanted to know why their client didn't get wait-listed. We can interpret this task as "How would the features $\{X_1, X_2, X_3\}$ contribute to unit x_a *not receiving outcome maybe and receiving no instead*, relative to the average outcome of other units drawn assuming marginally independent features with probability distributions defined using the values *of the units of the training population that received maybe*". In other words, we set $g = \mathbb{1}_{f(x)=\text{maybe}}(x)$, $\mathcal{Z} = \mathcal{X}$, and $p_S^{\mathcal{D}'}$ are created using the marginal fill-in method, and $\mathcal{D}'$ contains only the units from training set that received outcome *maybe*. This explanation is shown in Figure 2b.

Discussion. The two explanations are significantly different. In 2a, X_1 is the most important, and X_3, an indication of discrimination, is also high. In 2b, X_1 is not contributing, and X_3 is the main reason this unit did not receive a different outcome. Note that, depending on the definition of discrimination, the difference in the impact of the protected feature X_3 between the two explanations could affect whether this outcome is considered discriminatory.

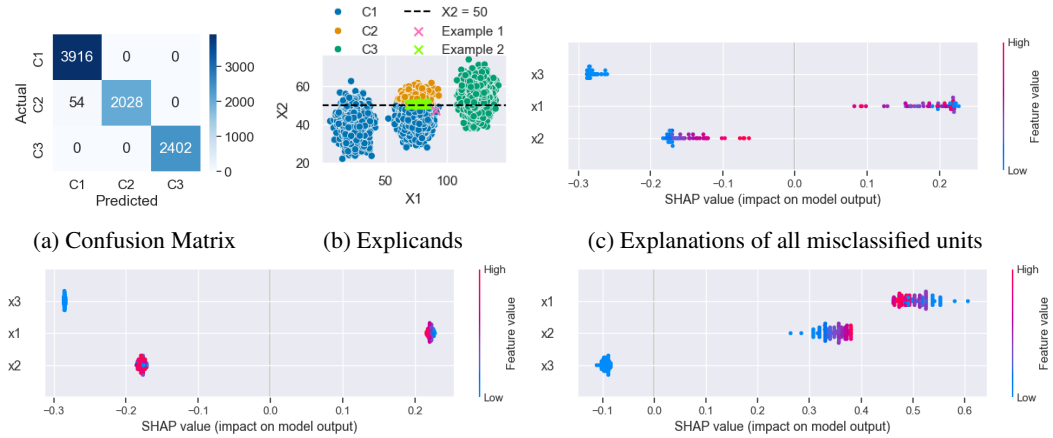
Validation. One could make the argument that the second explanation is answering the agent's question directly. Since we know exactly how outcomes are assigned for this dataset, we know that this unit received *no* because of rule 5. We also know that the question "why isn't the outcome *maybe* and is *no*" would require that the unit qualifies for either of the two rules 6 and 7. If the unit qualified for 6, then X_3 should be 0, and, if the unit qualified for 7, then X_2 should be more than 50, making X_2 and X_3 the only important features for the question, as shown in the second explanation.

4.2 Case study 2: Synthetic Dataset - Understanding Prediction Errors

We will examine a case where explainability can help us understand prediction errors. This example serves a dual role. In addition to arguing for the careful selection of the Shapley value parameters, we also point out that choosing the parameters carefully can expand their utility. We omit the Shapley card in the interest of space. The dataset and model are the same as in Case Study 1.

Context & Parameter Setting. Assume we want to understand why the items in Figure 3a are misclassified. We explain the missclassified units and the units that were correctly classified, and

[1] but with no sampling of the background data to remove the approximation error from the comparison



(a) Confusion Matrix (b) Explicands (c) Explanations of all misclassified units

(d) Explanations of units correctly label as "no" (e) Explanations of units correctly label as "maybe"

Figure 3: Model and Dataset information for case studies 1 & 2, and explanations for case study 2.

compare the explanations. In Figure 3, we show the explanations for the misclassified items that are classified as *no* instead of maybe, the explanations of similar items classified as *no* correctly, and similar items classified as *maybe* correctly. To identify similar items, we find the units that are correctly classified and have a minimum Euclidean distance from the incorrectly classified units, shown in Figure 3b. To create the explanations, we use $g = \mathbb{1}_{f(x)=\text{maybe}}(x)$ like in the previous example. The reason we want to use g for this example is to remove the effect of the feature importance of the units when they are classified in *Yes*. We provide the same explanations using $g = f$ in the Appendix Figure 8. There, we can see that even in this simple dataset, the importance of X_1 is fluctuating more; recall that all units considered have almost identical X_1 values.

Discussion. Looking at the explanations, we can see that the misclassified units, whose real label is *Maybe*, receive explanations almost identical to the units whose label is correctly classified as *No*. Additionally, the labels of the units that are correctly classified as *Maybe* mainly differ in the contribution of X_2. We conclude that the model is underestimating the effect of X_2 for these units.

Validation. Since we have access to the model, we can use this information to confirm that X_2 is the reason these units are misclassified. Indeed, changing only the value of X_2 can flip the label for these units at a threshold of about 50.3 instead of 50. While we could do this test without these explanations, the explanations provide us with a specific hypothesis that we can easily confirm.

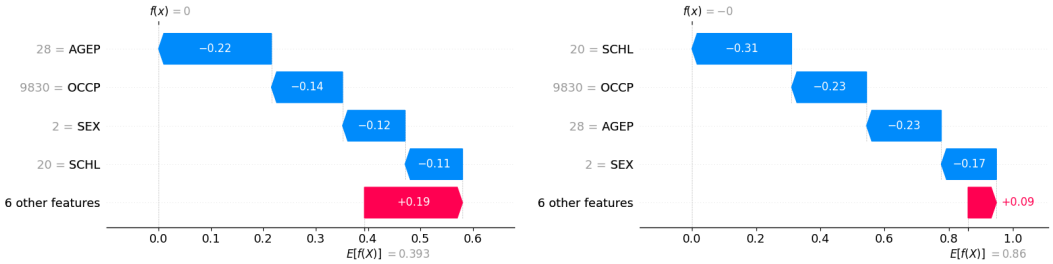
4.3 Case Study 3: Real Data - Average vs. Targeted Local Explanation

Dataset & Model. We use ACSIncome [17, 18] from CA (2018), and logistic regression. We train-test split 80%-20% and the accuracy is 0.8166. The task is to predict if income is more than 50K.

Context & Parameter Setting. We explain a 28-year old, female, full-time federal employee with an associate's degree. Her salary is below 50K, and we want to know how a higher degree would impact her salary. We produce two explanations, one using the entire population as $\mathcal{D}$ and one using only full-time federal employees that have associate's degrees or higher. The outcome is binary so $g = f$. This dataset contains many categorical variables that require one-hot encoding. We use $\mathcal{Z} \to \mathcal{X}$, so the explanation uses the input features, and the model uses one-hot encoded [2], see Section 2.2.

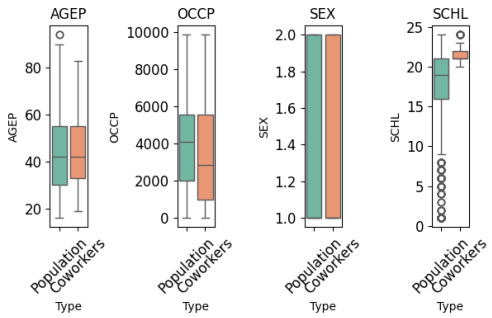
Discussion & Validation. In Fig. 4 we see that "SCHL" has a much higher importance in the second setting. In Fig 4c we can see how the means of the features shift when we subselect for this person's relevant colleagues. That is not surprising, real data generally has structure and subselecting the background data based on some feature-based criterion will affect the feature distributions of other features. If a feature whose mean changed was used by the model, then this subselection will also affect the means of the outcomes (see lemma 2.3), and if that feature is also important, then the shift of the model outcomes will be significant and the subselection will end up affecting the explanations.

[2]Note: in Python, we can implement the mapping $m : \mathcal{Z} \to \mathcal{X}$ using a scikit-learn [38] pipeline.



(a) Explanation when $\mathcal{D}$ is the population

(b) Explanation when $\mathcal{D}$ consists of similar people



(c) Shift of means for Case Study 3

Figure 4: Explanations for Case Study 3 on ACSIncome.

We see that in this case, the subselection of the relevant colleagues in combination with the model translates the shift of the mean of SCHL into a significantly higher feature importance. In short, we confirm that changing the background data based on the question can significantly impact the explanations [32, 48, 10]. See also Fig 9 in the Appendix for the complete explanations from Fig. 4.

4.4 Case Study 4: Image Data - Impact of Stochastic Subsampling

Dataset & Model. For this experiment, we used the MNIST image classification setup from the official SHAP documentation [3]. Were they utilize a CNN trained via PyTorch with the SHAP Deep Explainer, a marginal approach [30].

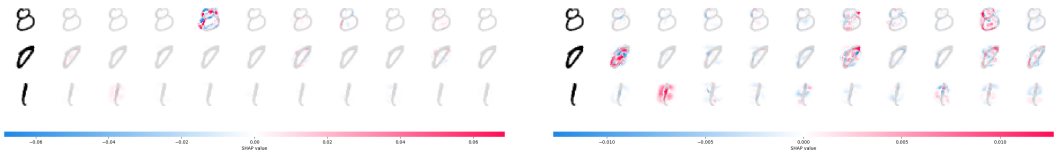
Context & Parameter Setting. This Case Study aims to illustrate how the multiplicity problem manifests in standard, real-world workflows. Following the workflow of this SHAP documentation experiment, we explained three specific digits using two different random subsets of background data ($\mathcal{D}$)—a parameter often subsampled in practice. In the following experiment, the model and the specific images being explained remain identical; only the background datasets (sampled from the held-out test set) are varied. The model and all parameters, including the size of the background dataset (100), are taken as is from the SHAP example. The context for this experiment can be seen as whether the standard out-of-the-box average local explanation differs when we use different subsets of background data sampled at random.

Discussion & Validation. As shown in Figure Fig. 5, the highlighted "important" pixels change significantly between the two runs. For instance, many pixels in the bottom-middle region appear as important only in the second set of explanations. This demonstrates that even common preprocessing choices, like random subsampling, introduce a form of "implementation noise" that can alter the explanation's narrative.

4.5 Case Study 5: Image Data - Differences Between Digits

Dataset & Model. Same as the previous Case Study.

[3] `https://shap.readthedocs.io/en/latest/example_notebooks/image_examples/image_classification/PyTorch%20Deep%20Explainer%20MNIST%20example.html`



(a) First random subset of background data

(b) Second random subset of background data

Figure 5: Explanations for Case Study 4. Shapley value explanations depend on the background dataset even when using random subsampling. In this plot, we build on an example from SHAP documentation. We show Shapley value explanations of three images from MNIST using a CNN and Deep Explainer, and different background data. Both explanations use the same model. This SHAP visualization shows the image being explained in the first column and explanations for classes 0-9 from left to right. We see that depending on the background data, different pixels are important for different classes.

(a) The background data consists of images of 1 or 7 only

(b) The background data consists of images of 5 or 8 only

Figure 6: Explanations for Case Study 5. Shapley value explanations can explain meaningful questions for image data. In this plot, we continue with the same setup as Figure 5.

Context & Parameter Setting. We further investigate how targeted parameter choices correspond to distinct semantic questions (Fig. 6). We explained the same three images of digits "1" and "7" under two different conditions:

- Intra-class comparison: Using a background of only "1" and "7" images. This answers: "What distinguishes these specific instances from the general 1/7 category?"
- Inter-class comparison: Using a background of "5" and "8" images. This answers: "What features distinguish a 1 or 7 from a 5 or 8?"

Discussion & Validation. The resulting importance maps are qualitatively different. In the first case, only pixels typical of 1s or 7s are highlighted. In the second, features shared with the 5s or 8s—such as the lower-left curve of an "8"—are highlighted as important to the explanation. These results confirm that SHAP outputs are not just sensitive to "noise," but are fundamentally driven by the user's implicit or explicit parameter choices. We believe this strongly supports our argument that documenting these parameters via a Shapley Card is essential for real-world interpretability.

5 Alternative Views

In this section, we address two credible counter-arguments to our proposal: the preference for standardization over parametrization, and the view that explanation multiplicity arises solely from a lack of causal identification.

View 1: Standardization is preferable to parametrization. While explanation multiplicity is widely acknowledged, a prevalent response in the literature is to resolve it through standardization—either by proposing a single "canonical" implementation or by aggregating multiple methods [5, 32, 45, 29]. This perspective effectively treats the Shapley value's axiomatic uniqueness as a mandate for a single universal configuration, typically defaulting to the marginal expectation over the training distribution. From this perspective, exposing multiple degrees of freedom $(g, \mathcal{D}, \mathcal{Z}, p_{\mathcal{S}})$ perhaps can be seen as imposing excessive cognitive load on stakeholders and hindering comparison across different models.

Response: We contend that standardization prioritizes consistency over validity, creating a risk of *misalignment*. As we demonstrated in Section 2, different parameter choices answer fundamentally

different questions. The game-theoretic guarantee of uniqueness holds only *for a fixed game*; it does not dictate which game is appropriate to play. By fixing a canonical configuration, one implicitly fixes a canonical question, which may be irrelevant to the stakeholder's actual context [9, 48]. We argue that the complexity of the Shapley Card is not artificial; it reflects the inherent complexity of explainability. Hiding these choices does not eliminate the normative assumptions; it merely obscures them, leading to a false sense of objectivity.

View 2: Multiplicity is a failure of causal identification. Some works suggest that explanation multiplicity is a side effect of lacking knowledge regarding the underlying data-generating process. Work by [26, 23] suggests that if the underlying Structural Causal Model (SCM) of the data were fully known, there would be a single, uniquely correct Shapley value based on do-calculus interventions. Under this view, our framework's parameters could be seen as simply proxies for missing causal knowledge.

Response: We agree that causal insight clarifies $p_{\mathcal{S}}$, but we maintain that it does not resolve multiplicity. Even with a perfect SCM, normative choices remain: one must still select the reference population $\mathcal{D}$ (e.g., relative to the population mean vs. a specific control group) and the profit function g (e.g., explaining the raw probability vs. the classification label) remain subjective definitions of the question, not objective properties of the model/data. Causal knowledge tells us how variables interact, but it cannot tell us which interaction is relevant to the stakeholder's specific question.

6 Summary

In this position paper, we provided a unifying Shapley value definition and demonstrated that feature importance is not an intrinsic property of a model or the data, but an outcome strictly defined by the parameters of the Shapley game. Since distinct configurations represent semantically distinct explanations, explanation multiplicity is expected; as a community it is our responsibility to determine which explanation aligns with a specific inquiry, or to acknowledge when no such alignment exists. To assist with this selection we issue a call to action, and to ensure trustworthiness, we propose Shapley Cards, arguing that reporting a value without its defining parameters obscures the underlying question and its formulation—rendering interpretation impossible. By shifting the conversation from "which method is best" to "which question is being asked," we believe the field can move toward a more transparent, reproducible, and scientifically rigorous science of explainability.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2020. URL `https://arxiv.org/abs/1903.10464`.

[2] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298: 103502, 2021.

[3] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International conference on machine learning*, pages 272–281. PMLR, 2019.

[4] John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317, 1964.

[5] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations, 2020. URL `https://arxiv.org/abs/2005.00631`.

[6] Dillon Bowen and Lyle Ungar. Generalized shap: Generating multiple types of explanations in machine learning, 2020. URL `https://arxiv.org/abs/2006.07155`.

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

[8] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2008.04.004. URL https://www.sciencedirect.com/science/article/pii/S0305054808000804. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[9] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data?, 2020. URL https://arxiv.org/abs/2006.16234.

[10] Hugh Chen, Scott M Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature communications*, 13(1):4512, 2022.

[11] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

[12] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

[13] Tanya Chowdhury, Yair Zick, and James Allan. Rankshap: Shapley value based feature attributions for learning to rank, 2024. URL https://arxiv.org/abs/2405.01848.

[14] Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *International conference on artificial intelligence and statistics*, pages 3457–3465. PMLR, 2021.

[15] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 203–214, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534159. URL https://doi.org/10.1145/3514094.3534159.

[16] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016. doi: 10.1109/SP.2016.42.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *CoRR*, abs/2108.04884, 2021. URL https://arxiv.org/abs/2108.04884.

[18] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. ACSIncome: Folktables US Census Data. https://github.com/socialfoundations/folktables, 2024.

[19] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. A study on the interpretability of neural retrieval models using deepshap. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1005–1008, 2019.

[20] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[21] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold, 2021. URL https://arxiv.org/abs/2006.01272.

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[23] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

[24] Maria Heuss, Maarten de Rijke, and Avishek Anand. Rankingshap – listwise feature attribution explanations for ranking models, 2024. URL https://arxiv.org/abs/2403.16085.

[25] Robert Hu, Siu Lun Chau, Jaime Ferrando Huertas, and Dino Sejdinovic. Explaining preferences with shapley values. *Advances in Neural Information Processing Systems*, 35:27664–27677, 2022.

[26] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[27] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. In *International conference on learning representations*, 2021.

[28] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kumar20e.html.

[29] Tony Liu and Lyle Ungar. Towards cotenable and causal shapley feature explanations. In *AAAI 2021 Workshop: Trustworthy AI for Healthcare*, 2021.

[30] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[31] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[32] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction*, pages 17–38. Springer, 2020.

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229. ACM, January 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

[34] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[35] Christopher Musco and R. Teal Witter. Provably accurate shapley value estimation via leverage score sampling, 2025. URL https://arxiv.org/abs/2410.01917.

[36] Lars HB Olsen, Ingrid K Glad, Martin Jullum, and Kjersti Aas. Using shapley values and variational autoencoders to explain predictive models with dependent mixed features. *Journal of machine learning research*, 23(213):1–51, 2022.

[37] Eliana Pastor, Luca de Alfaro, and Elena Baralis. Identifying biased subgroups in ranking and classification. *arXiv preprint arXiv:2108.07450*, 2021.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Venetia Pliatsika, Joao Fonseca, Kateryna Akhynko, Ivan Shevchenko, and Julia Stoyanovich. Sharp: Explaining rankings and preferences with shapley values. *Proceedings of the VLDB Endowment*, 18(11):4131–4143, 2025.

[40] Lloyd S Shapley et al. A value for n-person games. 1953.

[41] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[42] Helge Spieker, Jørn Eirik Betten, Arnaud Gotlieb, Nadjib Lazaar, and Nassim Belmecheri. Rashomon in the streets: Explanation ambiguity in scene understanding. In *Proceedings of the AAAI Symposium Series*, volume 7, pages 249–256, 2025.

[43] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

[44] Erik Štrumbelj, Igor Kononenko, and M Robnik Šikonja. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10):886–904, 2009.

[45] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[46] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR, 2020.

[47] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[48] Katarzyna Woźnica, Katarzyna Pekala, Hubert Baniecki, Wojciech Kretowicz, Elżbieta Sienkiewicz, and Przemysław Biecek. Do not explain without context: addressing the blind spot of model explanations. *arXiv preprint arXiv:2105.13787*, 2021.

A Dataset

In Figure 7, we illustrate the features and the outcomes of the dataset used in the examples for men and women respectively.

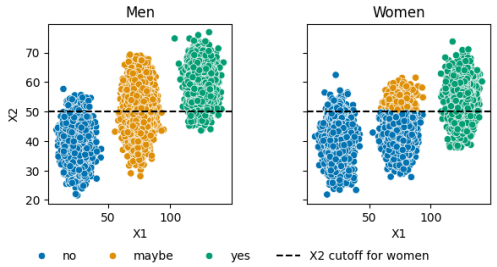


Figure 7: Feature values and outcomes for men ($X_3 = 0$) on the left and women ($X_3 = 1$) on the right. The dashed line indicates the cutoff for 'C2' for women.

B Case Study 2 - Supplementary

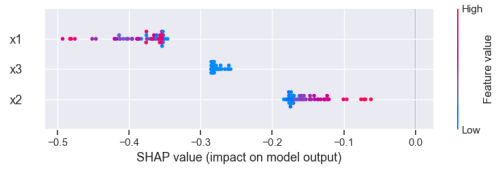
In Figure 8. We replicate the experiment of example 2 explaining f directly instead of $g = \mathbb{1}_{f(x)=\text{maybe}}(x)$. We include the comparison of these two experiments in the main text.

C Case Study 3 - Supplementary

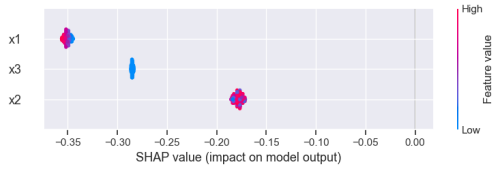
In Figure 9 we provide the complete explanation from Case Study 3.

D Case Study 6: Beyond Shapley values

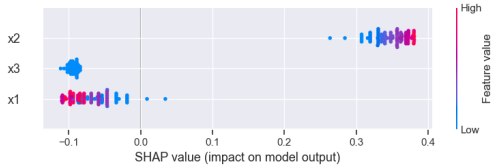
We run a simple experiment using Partial Dependence Plots (PDPs) [20] to illustrate that explanation multiplicity is not unique to Shapley values.



(a) Explanations of all missclassified units using $g = f$ as the profit function

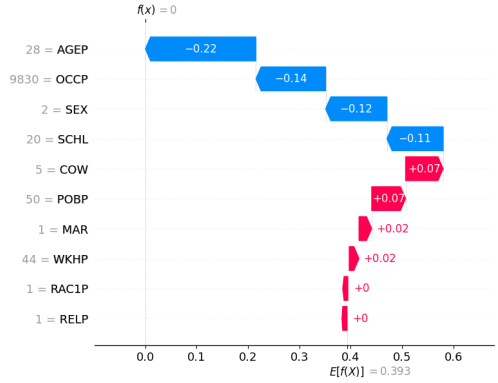


(b) Explanations of units correctly label as "no" using $g = f$ as the profit function

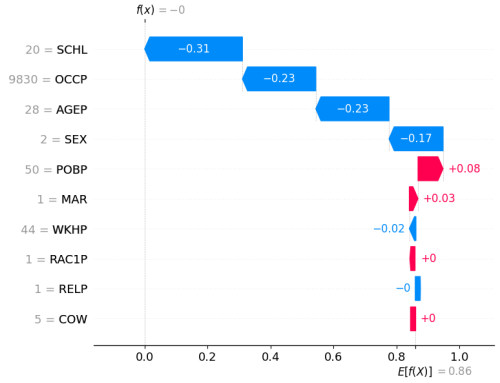


(c) Explanations of units correctly label as "maybe" using $g = f$ as the profit function

Figure 8: Explanations of different groups of units for the second example using $g = f$.



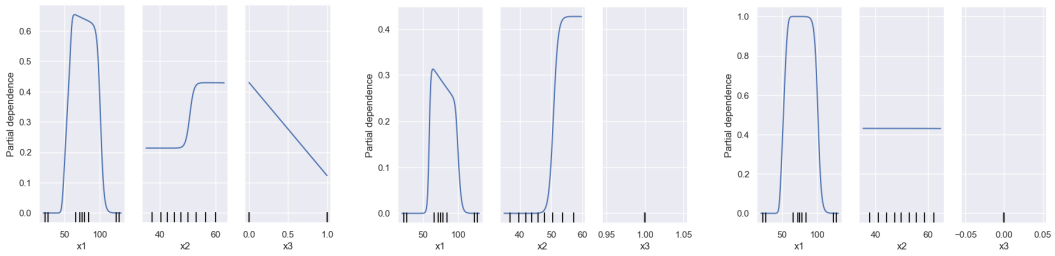
(a) Explanation when $\mathcal{D}$ is the population



(b) Explanation when $\mathcal{D}$ consists of similar people

Figure 9: Extended explanation of Case Study 3.

Partial Dependence Plots (PDP) also depend on the background data. In Fig. 10, we show PDPs for the survival show audition dataset introduced for Case Study 1, and the outcome class "Maybe" for three different background datasets: 10a for the entire population, 10b for women and 10c for men. In 10a, we see dependence on the gender (X_3). Looking at this dependence in more detail in the next two figures, we notice that for women, X_2 becomes important after the threshold of 50, correctly reflecting the fact that women need an X_2 value of 50 or higher to be classified in this class. This shows that the PDPs will produce different and meaningful results depending on the background data.



(a) Entire population (b) Women (c) Men

Figure 10: Different subsets of the background data result in PDPs with different functional form for the same model and data.