Explaining Models*

Kai Hao Yang† Nathan Yoder‡ Alexander K. Zentefis§

May 9, 2025

Abstract

We consider the problem of explaining models to a decision maker (DM) whose payoff depends on a state of the world described by inputs and outputs. A true model specifies the relationship between these inputs and outputs, but is not intelligible to the DM. Instead, the true model must be *explained* via a simpler model from a finite-dimensional set. If the DM maximizes their average payoff, then an explanation using ordinary least squares is as good as understanding the true model itself. However, if the DM maximizes their worst-case payoff, then *any* explanation is no better than no explanation at all. We discuss how these results apply to policy evaluation and explainable AI.

JEL classification: C50, D81
Keywords: models, decision-making, model explanations, least squares

*We thank Arjada Bardhi, Mira Frick, Paul Goldsmith-Pinkham, Kevin He, Ryota Iijima, Omer Tamuz, Akhil Vohra, Mark Whitmeyer, and conference participants at the Kansas Workshop in Economic Theory for very helpful comments.

†Yale School of Management, Email: kaihao.yang@yale.edu
‡University of Georgia, John Munro Godfrey Sr. Department of Economics, Email: nathan.yoder@uga.edu
§Hoover Institution, Stanford University, Email: zentefis@stanford.edu

1 Introduction

People must often make decisions in environments that are too complicated for them to understand. Policymakers evaluate social programs whose potential treatment effects are heterogeneous, highly nonlinear, or have spillovers. Regulators design rules for complex artificial intelligence models deployed in society without truly knowing how these models work. How useful to decision makers can intelligible *explanations* of their environments be instead?

In this paper, we study this question by considering the problem of a decision maker (henceforth DM) who encounters a model that is too complicated to understand, and instead must rely on an explanation of it. The DM's payoff depends on their action and the state of the world, where the latter is described by inputs and outputs. Inputs follow a known distribution, and a single *true model* specifies the relationship between inputs and outputs. For example, this true model could be the relevant data-generating process (DGP) that occurs in nature or the DGP that results from a complex artificial system, such as a large scale statistical or artificial intelligence (AI) model.

The key novel feature of our setting is that the space of true models is much larger than the space of *intelligible models* that the DM can understand. For example, the space of true models might contain all deep neural networks, but the space of intelligible models might contain only nth degree polynomials. For the DM to incorporate information about the true model into their choice of action, the true model must first be *explained* by mapping it to an intelligible model.

We focus on mappings between the space of true models and the space of intelligible models—what we call *explainers*—that have two consistency properties. First, if the true model is already intelligible, the explainer should not explain it with a different model. Second, the explainer should be linear, so that it preserves the structure of the space of possible models. Together, these criteria amount to the explainer being a *linear projection* of the true model onto the space of intelligible models. This class contains most tools used in practice to explain models, including ordinary least squares (which we consider formally in Section 3) and leading techniques for AI explainability (which we discuss in Section 2).

The paper's setting captures many situations in which decision makers confront complicated models that require an explanation. For instance, policymakers often evaluate social programs whose treatment effects (the outputs) depend on the demographic characteristics of the affected population (the inputs) through a complex relationship (the true model), and the policymakers must choose which programs to implement (the action). Similarly, regulators write rules on the deployment of complex AI models in society. Consider a state's transportation authority crafting safety standards for self-driving vehicles. Road, traffic, and weather conditions (the inputs) enter a deep neural network (the true model) that directs

the car's speed and navigation (the outputs). The regulator must decide the areas of the community, if any, the autonomous vehicles are allowed to operate (the action).

We consider two ways that the DM might evaluate their payoff. In the first, the DM maximizes the expectation of their payoff over the distribution of possible inputs. In the context of the program evaluation example, a policymaker behaving this way would care about the average treatment effect of a program. In the second, the DM puts weight only on the worst-case input. In the context of the self-driving cars, a regulator behaving this way would care only about the self-driving car's navigation (and the possibility of an accident) under road conditions that would lead to the worst possible consequences.

The main results of the paper show that these two ways to evaluate payoffs have sharply contrasting implications for the usefulness of model explanations as decision aids. If the DM cares about the average payoff across inputs, we show that it is possible to give a *robust explanation* that is *always perfect*, simply by explaining that model with the ordinary least squares (OLS) method (Theorems 1 and 2).[1] That is, the OLS explanation always allows the DM to make a decision that is *robustly* better *across all models* that are consistent with the explanation, and in fact *perfect* (in the sense that the DM is as well off as if they understood the true model itself).

An important caveat to this result is that it applies to an OLS explanation that is calculated using the same distribution of inputs (and therefore outputs) that is used to calculate the DM's expected payoff, rather than a distribution that differs due to sampling error. In other words, it applies to an OLS explainer that has access to the entire true model (as might be the case if it is an AI model), but not necessarily an OLS *estimator* that only has access to a *sample* from that true model (as might be the case if the model is a data-generating process found in nature). Theorem 3 shows that this distinction is not benign: The robust explanation offered by OLS is not robust to sampling error, even when that sampling error is small. In particular, there is always a sampling error such that OLS explanations cannot allow the DM to robustly improve their payoff *at all*. This has important implications for empirical work: Theorem 3 implies that OLS estimates can only be useful for decision making when combined with assumptions about the form of the relationship being estimated or about the way that sampling error occurs.

On the other hand, when the DM cares about their worst-case payoff, the prospects for explanation are grim no matter which explainer is used. Specifically, *any* explanation from *any* explainer is no better than having *no explanation at all* (Theorem 4). Intuitively, any explainer projects the infinite-dimensional space of possible true models onto a finite-dimensional space of explanations (i.e., the space of intelligible models). This limits the

[1]Here, an OLS-based explanation provides the coefficients from a linear regression of the outputs on the inputs.

information that can be recovered about the true model to a finite-dimensional sufficient statistic. Since there are infinitely many inputs, this statistic is not useful to a DM who cares about the worst-case input. In fact, this intuition for Theorem 4 extends to the intermediate case of an ambiguity-averse DM in the sense of Gilboa and Schmeidler (1989): If the DM's set of priors has higher dimension than the set of intelligible models, Theorem 5 shows that any explainer is unhelpful.

Outline The remainder of the paper proceeds as follows. Section 2 describes the paper's setting, and discusses its relationship to the literature on model evaluation (e.g., Fudenberg and Liang 2020). Section 3 and Section 4 provide the main results, the former when the DM cares about their average payoff, and the latter when they care about their worst-case payoff. Section 5 provides a discussion of the paper's findings. Section 6 describes the relationship between our work and several strands of related literature. Section 7 concludes.

2 Setting

Inputs and Outputs A state of the world is $(x, y) \in X \times Y$, where $X \subseteq \mathbb{R}^K$ is a convex set with $\dim(X) = K$, and Y is $\mathbb{R}^M$. For any state of the world $(x, y) \in X \times Y$, component $x \in X$ is interpreted as an *input* (or vector of exogenous variables) and component $y \in Y$ is interpreted as an *output* (or vector of endogenous variables). Inputs $x \in X$ follow a distribution μ_0.

Actions and Payoffs A decision maker (henceforth DM) chooses an action a from a finite set $A = \{a_1, \ldots, a_{|A|}\}$. The DM's payoff depends on the state of the world and the action chosen. Let $u : X \times Y \times A :\to \mathbb{R}$ denote the DM's payoff function. Throughout much of our analysis, we assume that u is *separable*, in the sense that $u(x, y, a) = w_0(a)+x\cdot w_1(a)+y\cdot w_2(a)$ for some functions $w_0 : A \to \mathbb{R}, w_1 : A \to \mathbb{R}^K, w_2 : A \to \mathbb{R}^M$.

True Models A *true model* is a bounded Borel measurable function $f : X \to Y$. Given an input value $x \in X$, a true model f specifies the relationship between inputs and outputs via $y = f(x)$.

Let $F \subseteq Y^X$ be the set of *possible* true models. Note that a true model could be highly complex: f, for instance, could be nonlinear, discontinuous, non-differentiable, a realization of a multi-dimensional Brownian path, or defined by a deep neural network.

Example 1 (Treatment Effects)**.** Consider a policymaker who chooses whether to implement a *treatment* $a \in \{0, 1\}$ in a population described by *covariate vectors* $x \in X$. Each output $y \in Y = \mathbb{R}^M = \mathbb{R}^2$ describes the potential outcomes of the treatment, so that $y_0 \in \mathbb{R}$ is the

outcome without treatment and y_1 is the outcome with treatment. The policymaker's payoff is

$$u(x, y, a) = y_a = y \cdot w(a)\,,$$

where $w(0) = (1, 0)$ and $w(1) = (0, 1)$. The outcome y_a under treatment a depends on the covariates x through a true model $f = (f_a)_{a \in A}$.

Example 2 (Self-Driving Car Regulation)**.** A regulator needs to set policies for self-driving cars by choosing among finitely many rules $a \in A$ (e.g., speed limits, number of approved licenses, areas to allow for self-driving). Inputs $X \subseteq \mathbb{R}^K$ denote all possible conditions surrounding a vehicle (e.g., lane markings, weather, infrastructure, traffic, visibility). An output is denoted by $y \in Y \subseteq \mathbb{R}^M = \mathbb{R}^{|A|}$, so that y_a is the expected net benefit of self-driving under rule a (taking into account potentially improved traffic efficiency and the possibility of accidents). The regulator's payoff is

$$u(x, y, a) = y_a - c(a) = y \cdot w_2(a)\,,$$

where $w_2(a_m)$ is a vector in $\mathbb{R}^{|A|}$ whose m-th component equals 1 and all other components equal zero, and $c(a)$ is the fixed cost of implementing rule a. The expected net benefit given rule a depends on condition x through a true model f, which is determined by the autonomous vehicle's algorithms, so that $f_a(x) = \mathbb{E}[y_a|x]$ is the expected net benefit when the condition is x and the rule is a.

Intelligible Models To capture the idea that the true model might be highly complicated and thus unintelligible to the DM, we consider a set Φ of *intelligible models*, where $\Phi \subseteq F$ is a finite-dimensional linear subspace that contains the constant function that always takes value of 1. Only models in Φ are intelligible to the DM, in the sense that the DM can only distinguish two different models, ϕ_1 and ϕ_2, if these models both belong to Φ. For instance, Φ could be the set of nth degree polynomials of x, which can be described by finitely many coefficients.

Decision Problem Henceforth, we refer to a *decision problem* by a tuple (A, u, Φ), where A is the (finite) set of available actions for the DM, $u : X \times Y \times A \to \mathbb{R}$ is the DM's payoff, and Φ is the set of intelligible models for the DM.

Explainers and Explanations For any decision problem (A, u, Φ), the true model f may not be intelligible to the DM. However, it can be explained to the DM through an *explainer*.

Definition 1. An *explainer* for the decision problem (A, u, Φ) is a linear idempotent operator $\Gamma : F \to F$ such that $\Gamma(F) = \Phi$.

An explainer Γ maps the true model f to an intelligible model $\Gamma(f) \in \Phi$ — an *explanation* that helps the DM to understand the true model indirectly. We focus on explainers that have two consistency properties (idempotency and linearity) possessed by most explainers used in practice, such as ordinary least squares (which is used extensively in applications like Example 1, and which we consider formally in Section 3) and leading techniques for AI explainability (which we discuss later in this section). Idempotency ensures that explanation is consistent with the truth whenever possible: if the true model f is intelligible, then it should not be explained using a different model. Linearity, on the other hand, ensures that explanation preserves the structure of the space of possible models. As we discuss below, in settings where the true model is a conditional expectation function (as may be the case in, e.g., Example 1), this amounts to consistency with the law of iterated expectations.

The explanation $\phi \in \Phi$ that is provided by an explainer Γ may not be precisely the same as the true model. But it does allow the DM to rule out all models that are not *consistent* with ϕ, i.e., all models that are not in

$$\Gamma^{-1}(\phi) := \{f \in F : \Gamma(f) = \phi\}.$$

Since the DM cannot discriminate between the models in this set, we focus on explanations that are *robustly useful* to him across all of them: that is, they allow the DM to improve his worst-case payoff across the models that are consistent with any given explanation.

The relationship between a true model f, an explainer Γ, and an explanation $\phi = \Gamma(f)$ is summarized in Figure 1 below.

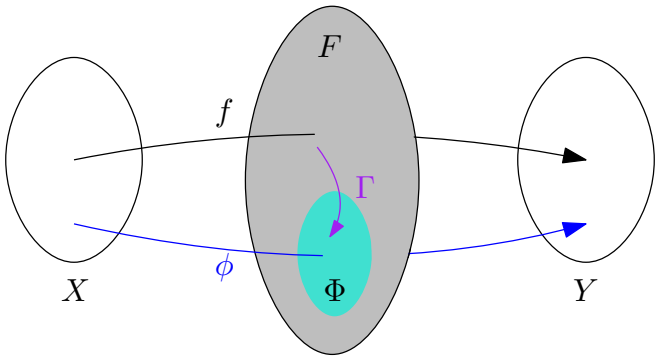

Figure 1: Models, explainers, and explanations. The figure depicts (1) the space F of possible true models f, which are functions from the space X of inputs to the space Y of outputs; (2) the subspace $\Phi \subset F$ of intelligible models ϕ; and (3) an explainer Γ that maps the space F of possible true models to the subspace Φ of intelligible models.

Discussion

Our setting and central question are both tightly linked to the literature on the evaluation of theoretical models (e.g., Fudenberg and Liang 2020; Montiel Olea, Ortoleva, Pai and Prat 2022; Fudenberg, Kleinberg, Liang and Mullainathan 2022; Andrews, Fudenberg, Liang and Wu 2023; Fudenberg, Gao and Liang 2024). Like those papers, we consider methods for simplifying a complex true relationship between inputs X and outputs Y using a map from the former to the latter — what they call a *prediction rule* and we call an explanation. We allow the method of simplification (the explainer) to vary, while they focus on mapping the truth to a prediction rule using minimum distance — what we call the least squares explainer. Instead, they vary the collection of maps from inputs to outputs that can be used to simplify the truth, which they call a model.[2] Our key departure is the following: In their settings, prediction rules are *approximations*, and their purpose is to approximate the truth as well as possible; in our setting, explanations are *information*, and their purpose is to be useful to a decision maker.

Our setup is also related to that considered in statistical decision theory (e.g., Wald 1949; Savage 1951). There, the statistician observes *data* from a *sampling distribution* that depends on the *state of the world*. Our baseline analysis abstracts from sampling error, and focuses on explanation instead: in our model, the DM observes an *explanation* from a deterministic *explainer* that takes the *true model* as its argument. As suggested by Wald (1949), we consider both *average* (Theorem 2) and *worst case* (Theorem 4) payoffs. However, a key difference in our setting is that the average or worst case is not only among models (which play the same role as "states" in statistical decision theory) but among their *inputs* (which have no counterpart in statistical decision theory).

When payoffs are separable, it ensures that for a decision maker who cares about their expected payoff — the case we consider in Theorems 1, 2, 3, and 5 — the only payoff-relevant characteristic of the model is its expected output. We note that separability does not necessarily require the map between inputs and outputs to be independent of the DM's action; this can be accomplished by letting each action correspond to part of the output vector, as in Example 1 and Example 2. Instead, it requires that for any given action, inputs and outputs enter payoffs independently.

As mentioned above, explainers used in practice are often linear. One especially relevant example is the ordinary least squares explainer considered in Section 3. Another is LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh and Guestrin 2016), which gives a local least-squares approximation to the true model.[3] A third is SHAP (SHapley Ad-

[2]We use the word "model" differently, to mean both the true relationship between inputs and outputs and an explanation.

[3]LIME also allows for penalizing the least-squares coefficient vector; it thus continues to satisfy property 2 (law of iterated expectations) when the penalty is the norm of the least-squares coefficient vector, as in

ditive exPlanations) (Lundberg and Lee 2017) which computes a linear approximation to the true model whose coefficients (on the coordinates x_i of the input vector) are a weighted average of the differences between the model's expected output conditional on x_S and conditional on $x_{S\cup\{i\}}$, for each $S \subseteq K \setminus \{i\}$.[4]

While we represent models as deterministic functions from inputs to outputs, our framework can also accommodate settings where the output is stochastic. One way to do this is to let one dimension x_K of the space of inputs represent a randomization device with distribution $\mu_0(\cdot|x_{-K})$. Alternatively, when only the average output $\mathbb{E}[y|x]$ is relevant to the DM's payoff at any given input x — i.e., when $u(x, y, a)$ is affine in y — we can simply interpret each possible true model as a conditional expectation function $f(x) = \mathbb{E}[y|x]$. Doing so provides an additional motivation for linearity: Then, linearity amounts to the consistency of explanations with the law of iterated expectations. That is, if a model is generated by randomizing between g and h — say, by using a state-independent randomization device $\varepsilon = \{g, h\}$ with $\mathbb{P}[\varepsilon = g] = \lambda$ — the explanation of that model should be the expected explanation of $g = E[y|\cdot, \varepsilon = g]$ and $h = E[y|\cdot, \varepsilon = h]$:

$$\Gamma(\lambda g + (1-\lambda)h) = \Gamma(\lambda \mathbb{E}[y \mid \cdot, \varepsilon = g] + (1-\lambda)\mathbb{E}[y \mid \cdot, \varepsilon = h]) = \lambda\Gamma(g) + (1-\lambda)\Gamma(h).$$

Or, put differently, the explainer Γ should not be affected by extra randomization devices that are not part of the state space $X \times Y$.

3 When are Explanations Robustly Useful?

We first explore a simple benchmark in which explaining a model to a DM has tremendous value. Specifically, we suppose that the DM cares about the *expectation* of his payoff under the distribution of inputs μ_0. This corresponds to situations in which decision makers are *utilitarians* and care only about the *average* performance of their actions. For example, policymakers may care only about the average treatment effect of an intervention; regulators or businesses may only care about the average performance of AI models they regulate or incorporate into their products. In what follows, we explore how explaining models can help the DM make better decisions when the DM only cares about the average. As a technical condition, we assume throughout this section that every possible true model $f \in F$ is square-integrable under the prior μ_0.[5] This allows for an well-behaved inner product structure on the space of models, and hence a well-defined orthogonal projection.

In particular, suppose the true model is f. Then such a decision maker obtains the

ridge regression, though not property 1 (consistency).

[4] As Lundberg and Lee (2017) show, this is equivalent to LIME with a quadratic loss function, a kernel based on the Shapley value, and no complexity penalty.

[5] That is, the set of all possible true models F is a linear subspace of $L^2(\mu_0)^M$.

expected payoff

$$U(f,a) := \mathbb{E}_{x\sim\mu_0}[u(x,f(x),a)]$$

when he takes action a. If he understood the true model, and chose the action that maximized this payoff, he would receive an expected payoff of

$$\overline{U}(f) := \max_{a\in A} U(f,a)\,.$$

By definition, $\overline{U}(f)$ is the highest payoff that the DM can achieve, given that the true model is f. Consequently, the performance of an explainer Γ in a given decision problem (A,u,Φ) can be evaluated by considering the gap between the benchmark full-information payoff $\overline{U}(f)$ and the DM's payoff $U(f,a(\Gamma(f)))$ when he takes an action $a(\Gamma(f))$ informed by the explanation $\Gamma(f)$.

But how should he choose such an action? Given any model $f \in F$ and any explainer Γ, there are many models that are consistent with the explanation $\phi = \Gamma(f)$. The DM is not able to identify which model $\hat{f} \in \Gamma^{-1}(\phi)$ is the true model given the explanation ϕ. Nonetheless, our first result shows that with the *ordinary least squares* explainer, this lack of identification is payoff-irrelevant for any DM who only cares about the average outcomes.

Definition 2. For any distribution $\mu \in \Delta(X)$, the *ordinary least squares (OLS) explainer for* μ is the unique orthogonal projection $\overline{\Gamma}_\mu$ from F onto Φ. That is, for each $f \in F$, $\overline{\Gamma}_\mu(f)$ is the unique element of Φ such that $\langle \phi, f - \overline{\Gamma}(f)\rangle_\mu = 0$ for all $\phi \in \Phi$, where $\langle\cdot,\cdot\rangle_\mu$ denotes the usual inner product in $L^2(\mu)^M$.[6] When μ is the true distribution of inputs μ_0, we denote this explainer $\overline{\Gamma} := \overline{\Gamma}_{\mu_0}$ and refer to it as *the OLS explainer*.

Theorem 1 shows that the OLS explainer perfectly identifies the payoff-relevant characteristics of the true model.

Theorem 1. *Suppose that (A,u,Φ) is a decision problem with separable payoffs. Then from the perspective of a utilitarian decision-maker, all models that are consistent with the same OLS explanation are payoff-equivalent. That is, for any action $a \in A$, and any models $f, \hat{f} \in F$ with $\overline{\Gamma}(f) = \overline{\Gamma}(\hat{f})$,*

$$U(\hat{f},a) = U(f,a)\,.$$

Theorem 1 shows that when the model is explained using the OLS explainer $\overline{\Gamma}$, even though there are many models that might be consistent with an explanation, and the DM

[6] Specifically, given $\mu \in \Delta(X)$, for any $f, g \in F$, define the inner product

$$\langle f,g\rangle_\mu := \mathbb{E}_{x\sim\mu}\left[\sum_{j=1}^{M} f_j(x)g_j(x)\right].$$

cannot identify which one is the true model, the DM's expected payoff when he takes a given action is the same across all these models. As a result, explaining models using the OLS explainer is always enough for the DM to identify all they need to make a decision. For intuition, recall that when payoffs are separable, an expected utility maximizer only cares about the model's expected output. OLS preserves that expectation because it is the orthogonal projection with respect to the inner product that uses the DM's prior μ_0.

A straightforward consequence of Theorem 1 is that OLS gives a *robust explanation* that is *always perfect*: it *always* (for any possible true model) gives an explanation that allows the DM to make a decision that is *robustly* (across all models that are consistent with the explanation) better, and in fact *perfect* (in the sense that the DM could do no better by knowing the true model).

Corollary 1 (Always Perfect Robust Explanations)**.** *Suppose that (A, u, Φ) is a decision problem with separable payoffs. Then for any true model f, the explanation $\phi = \overline{\Gamma}(f)$ robustly explains f:*

$$\inf_{g \in \overline{\Gamma}^{-1}(\phi)} U(g, a) = U(f, a). \tag{1}$$

In particular, the decision-maker can do no better than if she knew the true model:

$$\max_{a \in A} \inf_{g \in \overline{\Gamma}^{-1}(\phi)} U(g, a) = \overline{U}(f).$$

In fact, a utilitarian DM need not be sophisticated enough to compute the worst case described in (1) in order to get the full value of robust explanation from OLS. Instead, they can achieve their first-best value $\overline{U}(f)$ by naïvely choosing an action as if the explanation $\phi = \overline{\Gamma}(f)$ was the true model.

Theorem 2 (OLS is All You Need)**.** *For any decision problem (A, u, Φ) with separable payoffs, and for any true model $f \in F$, a utilitarian DM's first-best value can be achieved by treating the explanation as the true model under the ordinary least squares explainer $\overline{\Gamma}$. That is, for each explanation $\phi \in \Phi$, let $a^*(\phi) \in \arg\max_{a \in A} U(\phi, a)$. Then for all $\phi \in \Phi$ and all $f \in \overline{\Gamma}^{-1}(\phi)$,*

$$U(f, a^*(\phi)) = \overline{U}(f)\,.$$

Theorem 2 shows that even a DM that takes the explanation as literally true can achieve exactly the first-best benchmark via the ordinary-least square explainer. In other words, even when the set of intelligible models is very limited, explaining any complex model through OLS allows the DM to get the same payoff they would get if they understood the true model.

We note that Theorems 1 and 2 both rely on the assumption of separable payoffs. That is, for any given action, inputs and outputs must enter the DM's payoff independently. While

this assumption is appropriate in some settings (e.g., Examples 1 and 2), it is strong, and without it, Theorems 1 and 2 both fail. As it turns out, this foreshadows our results in the rest of the paper, which show that while explanation can be robustly useful (Theorems 1 and 2), that usefulness is fragile in several different ways.

Sampling Error

Theorems 1 and 2 suggest that explaining models using least squares can be very useful for a utilitarian decision maker who is concerned with their *average* payoff over all inputs x. This, however, relies on the fact that the distribution used to construct the OLS explainer is precisely the distribution μ_0 that is relevant for the DM's expected payoff. But in the real world, OLS is most often used as an *estimator* rather than an *explainer*: that is, rather than computing the least-squares approximation under the *population* distribution μ_0, practitioners compute it under a *sampling* distribution μ over the space of exogenous variables (i.e., model inputs). This may be due to lack of data (e.g., when f maps demographic variables to treatment effects, as in Example 1) or to economize on computational resources (e.g., when f is an AI model, as in Example 2).

Theorem 3 shows that the robust explanation offered by the OLS explainer is not robust to even small differences in these distributions — i.e., to even small *sampling error*.

Theorem 3 (Robust Explanations Are Not Robust to Model Uncertainty)**.** *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$, and that (A, u, Φ) is a decision problem with separable payoffs and a product set of intelligible models. The robust usefulness of least squares explanations is not robust to even small sampling error: For any open $\mathcal{M} \subseteq \Delta(X)$ with $\mu_0 \in \mathcal{M}$, there exists $\mu \in \mathcal{M}$ such that for any action a and explanation $\phi \in \Phi$,*

$$\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi)} U(f, a) = \inf_{f \in F} U(f, a).$$

In particular, the decision-maker can do no better than naïvely maximizing her expected payoff over all outputs:

$$\max_{a \in A} \inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi)} U(f, a) = \max_{a \in A} \inf_{f \in F} U(f, a) = \max_{a \in A} \inf_{y \in Y} \mathbb{E}_{x \sim \mu_0}[u(x, y, a)]$$

Intuitively, the robust usefulness of the OLS explanation $\phi = \overline{\Gamma}(f)$ relies on every possible true model $\hat{f}$ in its preimage $\overline{\Gamma}^{-1}(\phi)$ producing the same expected output $\mathbb{E}_{x \sim \mu_0}[\hat{f}(x)]$. Or, put differently, it relies on every element of the OLS explainer's kernel $\ker \overline{\Gamma} := \{g \in F | \overline{\Gamma}(g) = 0\}$ having an expected output of zero. In the appendix, Proposition 2 shows that there are distributions μ arbitrarily close to μ_0 such that for any action a, the kernel of the OLS estimator $\overline{\Gamma}_\mu$ contains elements $g^{w_2(a)}$ such that $g^{w_2(a)} \cdot w_2(a)$ has nonzero expectation. Since

$\ker \overline{\Gamma}_\mu$ is a subspace of F, it must also contain elements such that $\mathbb{E}_{x\sim\mu_0}[g(x)] \cdot w_2(a)$ — and hence $U(\phi + g, a)$ — is arbitrarily small.

4 When are Explanations Not Robustly Useful?

When the space of intelligible models is rich enough (but still finite dimensional), Corollary 1 shows that explanation with ordinary least squares robustly (i.e., across all models consistent with the explanation) achieves *expected* payoffs indistinguishable from those obtained with complete knowledge of the true model. In fact, Theorem 2 shows that treating the OLS explanation as if it were the true model yields the same expected payoff $\overline{U}(f)$ as having direct knowledge of the true model itself.

But models must often be explained to agents who want to make a decision that is robust across *inputs*, not just across models. Policymakers may want to ensure that an intervention has beneficial effects to *all* members of a population, not just on average (i.e., if the policymakers have a *Rawlsian* social welfare function). Likewise, regulators or firms may be most concerned about the most catastrophic effects that could result from adopting an AI model, not just the model's average performance.

To understand how explanations could benefit such a decision maker, suppose that he observes an explanation ϕ from an explainer Γ. Then for each action $a \in A$, the worst case payoff that is consistent with that explanation is

$$R(\phi, a|\Gamma) := \inf_{\substack{f\in\Gamma^{-1}(\phi)\\ x\in X}} u(x, f(x), a)\,.$$

In contrast, in the absence of an explanation, the DM's worst-case payoff from taking action a is

$$\underline{R}(a) := \inf_{\substack{f\in F\\ x\in X}} u(x, f(x), a)\,.$$

If such a DM benefits from receiving an explanation, it must be because it causes him to change his action; that is, because there is some pair of actions a, a' such that $\underline{R}(a') \geq \underline{R}(a)$ but $R(\phi, a|\Gamma) > R(\phi, a'|\Gamma)$. Unfortunately, in stark contrast to Theorem 2, Theorem 4 reveals that this is impossible: when the space of possible true models is rich enough, explanation cannot change the worst-case payoff from any action. Thus, explaining the true model is never helpful for making decisions that are robust across both models and inputs. In fact, this result extends beyond decision problems with separable payoffs: All that is needed is that payoffs from any given action *depend on one dimension of output*, i.e., we can write $u(x, y, a) = v(x, y \cdot w(a), a)$ for some $w : A \to Y$ and $v : X \times Y \times A \to \mathbb{R}$.

Theorem 4 (Worst-Case Model Outcomes are Inexplicable)**.** *Suppose that F contains all*

bounded Borel measurable functions $f : X \to Y$, and that (A, u, Φ) is a decision problem with payoffs that depend on one dimension of output for any given action. No explainer can provide an explanation whose usefulness is robust across both inputs and models: For any action a, any explainer Γ, and any $\phi \in \Phi$,

$$R(\phi, a|\Gamma) = \underline{R}(a).$$

In particular, the decision-maker can do no better than naively maximizing her worst case payoff over all states of the world:

$$\max_{a \in A} R(\phi, a|\Gamma) = \max_{a \in A} \underline{R}(a) = \max_{a \in A} \inf_{\substack{y \in Y \\ x \in X}} u(x, y, a). \tag{2}$$

Intuitively, the space of possible explanations is finite-dimensional, but the space of possible models is infinite-dimensional. The only way that a linear explainer can map from the latter to the former is by discarding information about all but finitely many of those dimensions (i.e., the true model's output values at all but finitely many input values).

In particular, suppose the DM observes an explanation ϕ^*. Proposition 1 below shows that for every possible value z of the dimension of output $y \cdot w(a)$ on which her payoff from a depends, and almost every possible input x, there is some model f with $f(x) \cdot w(a) = z$ that is consistent with that explanation. Since the DM's payoff is continuous and the space of inputs is convex, this is enough to ensure that the explanation does not change the infimum in (2).

Proposition 1. *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$. Let Γ be an explainer; let $\phi^* \in \Phi$ be an explanation; let $w \in Y \setminus \{0\}$ be a vector; let $z \in \mathbb{R}$. For all but finitely many $x \in X$, there exists $f \in \Gamma^{-1}(\phi^*)$ such that $f(x) \cdot w = z$.*

Note that the impossibility result of Theorem 4 does not reverse as the dimension of Φ increases. In other words, no matter how many models are intelligible, *any* explanation from *any* explainer provides no information that is useful in making decisions that are robust across both inputs and models. For instance, if Φ is the set of nth degree polynomials, then no matter how large n is, no explainer improves the payoff of a DM who cares about the worst case, because the worst-case payoff consistent with an explanation is approximately the same, no matter the explanation.

Together, Theorem 4 and Proposition 1 reveal that explanations of complicated models offer no assistance to a DM who wants to maximize her worst-case payoff, no matter how rich the set of (finite-dimensional) intelligible models is, and even without an assumption of separable payoffs. When the model is explained using OLS, even if the set of possible true models that are consistent with a given explanation is infinite-dimensional, they are all the

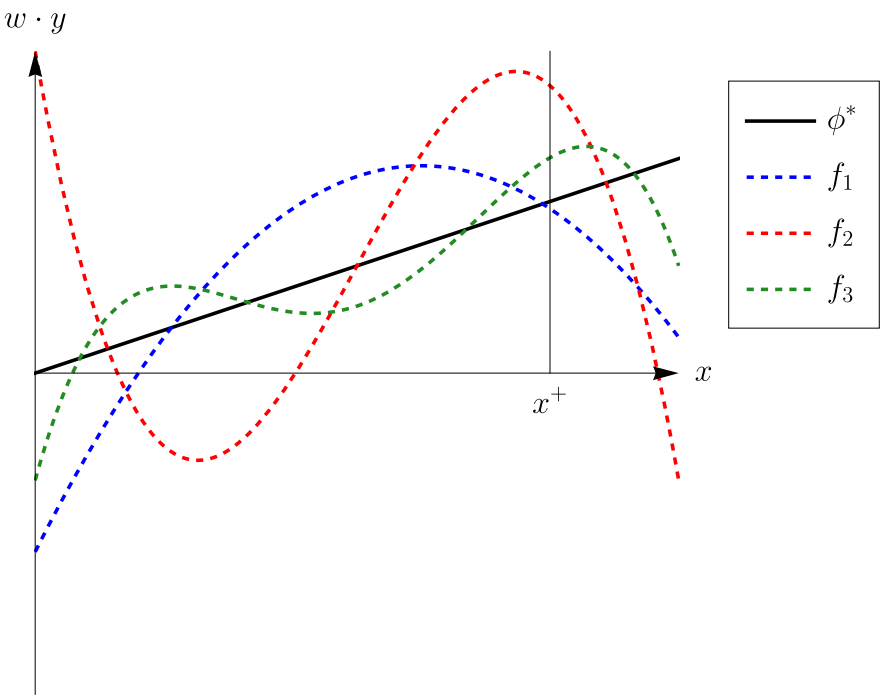

Figure 2: True models consistent with the same explanation. Proposition 1 shows that given an explainer Γ, at all but finitely many values x of the input and any explanation ϕ^*, one can take any value z of a dimension of the output $w \cdot y$ and find a model that is consistent with ϕ^* such that $w \cdot f(x) = z$. Figure 2 illustrates an example where $X = [0, 1]$, Γ is the OLS explainer $\overline{\Gamma}$, and μ_0 is uniform: At x^+, the models f_1, f_2, and f_3 give very different values of $w \cdot y$. But each is consistent with the explanation ϕ^*: $\overline{\Gamma}(f_1) = \overline{\Gamma}(f_2) = \overline{\Gamma}(f_3) = \phi^*$.

same *on average*. Hence, if the DM cares about the expected payoff, an OLS explanation is just as good as understanding the true model. But when the DM cares about their worst case payoff, the average model output is irrelevant. Rather, the set of payoffs that these models give the DM *at the worst-case inputs* determines the performance of an explainer. As Theorem 4 shows, the finite-dimensional nature of the set of intelligible models makes this set unaffected by an explanation.

To illustrate the implications of Theorem 4 and Proposition 1, we can revisit the treatment effect example of Example 1, but now with a Rawlsian DM concerned with the *worst possible* treatment effects. Suppose once more that the DM can understand explanations of the data generating process as an nth degree polynomial, but that any true model outside that class is unintelligible. Reporting the coefficients from a linear regression—which is standard practice in the treatment effects literature—is then intelligible to the DM, but will never alter their decisions. In fact, there is no explainer that can help the DM make a program evaluation when they care about those in the population who would be most disadvantaged by the policy.

Ambiguity Aversion

In Section 3, we considered the case of an expected utility maximizer who knows the distribution of inputs μ_0. What if the DM was instead ambiguity-averse in the sense of Gilboa and Schmeidler (1989), had a set of priors over inputs, and maximized the worst expected payoff over that set? Or, what if the DM does not know the true distribution of inputs μ_0 exactly, and wants to make a decision that is robust across models *and also robust* across a set of possible distributions $\mathcal{M}$?

Formally, suppose that an ambiguity averter observes an explanation from an explainer Γ. Then the worst-case payoff from action $a \in A$ that is consistent with his set of priors $\mathcal{M}$ and the explanation ϕ he observes is given by

$$R_{\mathcal{M}}(\phi, a|\Gamma) := \inf_{\substack{f \in \Gamma^{-1}(\phi) \\ \mu \in \mathcal{M}}} \mathbb{E}_{x\sim\mu}[u(x, f(x), a)]\,,$$

while without an explanation, his worst-case payoff is

$$\underline{R}_{\mathcal{M}}(a) := \inf_{\substack{f \in F \\ \mu \in \mathcal{M}}} \mathbb{E}_{x\sim\mu}[u(x, f(x), a)]\,.$$

Ambiguity aversion is a natural intermediate case between expected utility maximization (where Theorem 2 shows that an OLS explanation allows the DM to achieve the full-information payoff) and worst-case analysis (where Theorem 4 shows that explanation cannot improve the DM's payoff). Hence, we might expect the efficacy of explanation to be intermediate between those two cases as well. Unfortunately, it is not: Theorem 5 shows that when the set of possible distributions is higher-dimensional than the space Φ of intelligible models, robust explanation is impossible—even when, as in Section 3, the DM has separable payoffs, and thus cares only about the model's expected inputs and outputs.

Theorem 5 (Impossibility of Robust Explanation with Ambiguity Aversion)**.** *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$, and that (A, u, Φ) is a decision problem with separable payoffs. If an ambiguity averse DM has a set of priors with sufficiently high dimension, robustly useful explanation is impossible: For any action a, any explainer Γ, any convex $\mathcal{M} \subseteq \Delta(X)$ with $\dim(\mathcal{M}) > \dim(\Phi)$, and any $\phi \in \Phi$,*

$$R_{\mathcal{M}}(\phi, a|\Gamma) = \underline{R}_{\mathcal{M}}(a).$$

In particular, the decision-maker can do no better than naively maximizing her worst case

payoff over all outputs:

$$\max_{a\in A} R_{\mathcal{M}}(\phi, a|\Gamma) = \max_{a\in A} \underline{R}_{\mathcal{M}}(a) = \max_{a\in A} \inf_{\substack{y\in Y \\ \mu\in\mathcal{M}}} \mathbb{E}_{x\sim\mu}[u(x, y, a)]. \tag{3}$$

Theorem 5 shows that explanations of the true model are not robustly helpful to an ambiguity averse DM. Intuitively, just like in Theorem 4, the set of features of the true model over which the DM takes the worst case, $\{\mathbb{E}_{x\sim\mu}[f(x)]\}_{\mu\in\mathcal{M}}$, has higher dimension than the space of intelligible models. Thus, any explanation can only provide information about a lower-dimensional manifold in $\{\mathbb{E}_{x\sim\mu}[f(x)]\}_{\mu\in\mathcal{M}}$, and must provide no information about that manifold's complement—which is dense in $\{\mathbb{E}_{x\sim\mu}[f(x)]\}_{\mu\in\mathcal{M}}$.

Theorem 5 also provides a pessimistic perspective on Theorem 2 that complements that of Theorem 3. Each result shows, in slightly different ways, that Theorem 2 relies on the distribution of inputs used to compute the DM's payoff *exactly matching* the distribution of inputs used to compute the model's OLS explanation. To show that robust explanation is not robust to sampling error, Theorem 3 perturbs the former; to show that it is not robust to ambiguity about the distribution of inputs, Theorem 5 perturbs the latter.

5 Discussion

5.1 The Effectiveness of Explanations

Theorem 2 and Theorem 4 present a fundamental dichotomy between the two regimes when explaining complicated models. When a decision maker cares about her average payoff across inputs, it is possible to offer an explanation that is robustly first-best across models using the canonical OLS approach. But no explainer can offer a useful explanation that is robust across both models and inputs (Theorem 4) or even across both models and distributions of inputs (Theorem 5). Moreover, the robust explanation offered by OLS is not robust to sampling error (Theorem 3).

In the context of policymaking, Theorem 2 suggests that standard regression analyses are useful and powerful tools for summarizing and approximating the relationship between inputs and outputs for a utilitarian policymaker who cares about the average outcome. However, Theorem 4 suggests that when the policymaker wants to make a decision that is robust across *all* inputs, it is impossible for any regression analyses to provide useful guidance, unless some possible true models are ruled out *a priori*. As a result, *any* attempt at explaining the complicated data generating processes that occur in nature is then futile, as there are no explainers that can improve—even slightly—the policymaker's decisions.

Likewise, in the context of AI regulation, explaining a black-box AI model to a regulator could be extremely helpful to a regulator who wishes to improve average outcomes.

Nonetheless, it is impossible to enable better decisions about worst-case scenarios by explaining black-box AI models.

Together, our results suggest that the effectiveness of model explanations depends crucially on how the decision maker to whom the model is explained evaluates their payoff. In particular, in environments where the decision maker is concerned about the worst-case scenario, the availability of explanations of the true model—however sophisticated they are—do not alleviate those concerns.

5.2 The Value of Theory in Explanations

Our results that rule out the possibility of robust explanation (Theorems 4 and 5) each rely on a richness condition on the space of possible models: F contains all bounded measurable functions from the space of inputs to the space of outputs. Or, put differently, *no well-behaved function can be ruled out as a possible true model.*

But if some of these models can be ruled out as inconsistent with theory, then Theorems 4 and 5 do not apply, and it may be possible to offer a useful explanation that is robust across both models and inputs. For instance, if theory predicts that the effect of a treatment must be nondecreasing in a demographic variable, it may be possible to explain the true model in a way that is useful for a policymaker, even when that policymaker is Rawlsian and cares about the effect on those most disadvantaged by the treatment he chooses.

Thus, the message of our results is more nuanced than it might appear: it is not that explanations to a decision maker who cares about the worst case are *never* robustly useful, but that theory is *necessary* for such explanations to be robustly useful.

5.3 The Robustness of Reduced Form Estimation

Many empirical analyses aim to be as agnostic as possible about the way that the variables that they study are related. Instead of making "structural" assumptions about the functional form of that relationship, they provide a "reduced form" estimate of a linear approximation to it. For instance, if they wish to estimate the effect of a treatment á la Example 2, they might not make assumptions about the way that the treatment effect might depend on individual characteristics (i.e., limitations on the space of true models F, or a prior over F) and instead focus on estimating an average treatment effect (i.e., the expectation $\mathbb{E}_{x\sim\mu_0}[f(x)]$). Similarly, they might appeal to the central limit theorem to avoid making assumptions about the distribution of sampling error.

Our results give a sharp characterization of the usefulness of this approach to decision makers who rely on the results of such analyses. When there is no sampling error, and decision makers only care about the relationship between exogenous and endogenous variables *on average*, Theorem 2 shows that structural assumptions are unnecessary. But when sampling

error is present, Theorem 3 shows that this is no longer true: Least-squares estimates can only be useful for decision making when combined with assumptions about the model's functional form and/or the form of the sampling error. Likewise, when a decision maker is interested in a worst-case rather than average treatment effect, Theorem 4 shows that even in the absence of sampling error, least squares estimates are only useful when combined with structural assumptions.

We emphasize that this characterization does not apply to nonparametric estimation: While parametric estimators like OLS project the true model onto a finite-dimensional space of models Φ that can be easily interpreted, nonparametric estimators such as kernel density sacrifice interpretability by projecting the true model onto an infinite-dimensional space. Since the difference between the dimensionality of F and Φ is key to Theorems 3-5, such estimates can be useful for any DM even in the absence of structural assumptions: For instance, in the absence of sampling error, the kernel density estimate of the true model f is just f.

5.4 Recommendations vs. Explanations

Explanations are not robustly useful in the contexts considered by Theorems 4 and 5 because the space of intelligible models is finite-dimensional, but the space of true models is infinite-dimensional. However, the DM only cares about the model insofar as it helps them choose an action, and the set of actions is finite. This suggests a remedy to the negative results of Theorems 4 and 5: Instead of offering *explanations* (i.e., intelligible models that represent the true model), offer *recommendations* (i.e., inform the DM of the optimal action under the true model). That is, instead of using an explainer $\Gamma : F \to \Phi$, one should use a *recommender* defined by[7]

$$\begin{aligned} G : F &\to A \\ f &\mapsto \operatorname*{argmax}_{a \in A} \inf_{x \in X} u(x, f(x), a) \end{aligned}$$

Clearly, a recommender always makes the DM as well off as if he understood the true model. Moreover, unlike an explainer, a recommender places no cognitive demands on the DM. Instead of considering all possible true models that could produce an explanation, and evaluating the worst-case payoff for each action, the DM can simply follow the recommended

[7] Or in the ambiguity-averse case,

$$\begin{aligned} G_{\mathcal{M}} : F &\to A \\ f &\mapsto \operatorname*{argmax}_{a \in A} \inf_{\mu \in \mathcal{M}} \mathbb{E}_{x \sim \mu}[u(x, f(x), a)]. \end{aligned}$$

action.

However, a recommendation can only be successful if the decision maker's payoff can be incorporated into the recommender's design. If the same information about the true model must be used by many decision makers with heterogeneous preferences or even one decision maker with private information, a recommender may not deliver the full-information payoff, because it may not always be optimal for the decision maker(s) to follow the recommendation. Indeed, there is ample empirical evidence of people overriding model recommendations to make high-stakes decisions in several sectors of society like criminal justice, medicine, and finance (De-Arteaga, Fogliato and Chouldechova 2020; Jussupow, Benbasat and Heinzl 2020; Ludwig and Mullainathan 2021; Angelova, Dobbie and Yang 2023).[8]

6 Related Literature

Our paper sits adjacent to a large recent literature that considers the use of (potentially misspecified) models in decision making. Like us, these papers also consider environments where agents may be constrained in their ability to understand a payoff-relevant mapping. This may take the form of, for instance, a mapping from actions to outputs (e.g., Esponda and Pouzo 2016; Fudenberg, Lanzani and Strack 2021); a mapping from states to signal distributions (e.g., Schwartzstein and Sunderam 2021); or a mapping from states and past actions to signal distributions (e.g., Bohren and Hauser 2024). Of these, our paper is closest to the strand of the literature focusing on the evaluation of theoretical models (e.g., Fudenberg and Liang 2020, Montiel Olea et al. 2022), in which the relevant mapping is between inputs and (distributions over) outputs; see the Discussion in Section 2 for a detailed comparison of our settings and research questions.

As we note in Example 1, our paper also has implications for policy evaluation that connect to the literature on the optimal design of RCTs in a potential-outcomes setting (e.g., Banerjee, Chassang and Snowberg 2017; Chassang, Padró i Miquel and Snowberg 2012; Kasy et al. 2013; Banerjee, Chassang, Montero and Snowberg 2020; Chassang and Kapon 2022). Our focus is complementary to that taken by these papers: Rather than focusing on distributing measurement error across different dimensions of output (i.e., different potential outcomes), we focus on the consequences of the decision maker's inability to comprehend the true relationship between individual characteristics (inputs) and treatment effects.

Finally, there is a large literature in computer science on the use of explanations to help humans use predictions made by AI models to make decisions (e.g., Lai and Tan 2019; Bansal, Wu, Zhou, Fok, Nushi, Kamar, Ribeiro and Weld 2021). This focus is subtly different than ours: While we also consider the explanation of AI models that map inputs (e.g., features) to

[8] Iakovlev and Liang (2023) theoretically compare and contrast the important issue of choosing between human evaluators who use context to make predictions and algorithms that do not.

outputs (e.g., labels), in our setting the model is directly relevant to the DM's payoff, rather than providing suggestions that inform the DM about another data generating process. The explanations in our paper thus provide information about the underlying payoff-relevant relationship directly, instead of indirectly (by providing information about an approximating model).[9]

7 Conclusion

We consider the problem of explaining models to a decision maker (DM). The DM has a payoff that depends on their actions and the state of the world, where the latter is described by inputs and outputs. A true model specifies the relation between these inputs and outputs, but is not intelligible to the DM. For the DM to make a choice, the true model instead has to be *explained* using an intelligible model that belongs to a finite dimensional space. We show that if the DM maximizes their average payoff across inputs, then an explanation using ordinary least squares is arbitrarily close to as good as understanding the true model itself. However, if the DM maximizes their worst-case payoff across inputs, then *any* explanation offers no advantage over no explanation at all.

The paper's environment leaves room for continuing work. For example, we focus on a single decision maker, but a second agent could be introduced, one who provides explanations of models that may misalign with the interests of the decision maker.[10]

References

Andrews, I., D. Fudenberg, A. Liang, and C. Wu (2023) "The Transfer Performance of Economic Models," Working paper.

Angelova, V., W. S. Dobbie, and C. Yang (2023) "Algorithmic recommendations and human discretion," Working paper.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020) "A Theory of Experimenters: Robustness, Randomization, and Balance," *American Economic Review*, 110 (4), 1206–1230.

Banerjee, A. V., S. Chassang, and E. Snowberg (2017) "Decision theoretic approaches to

[9]In addition, instead of choosing an action after seeing an input and an explanation of the model at that input, the DM chooses a single, fixed action. These factors mean that our setup fits a different set of applications: For example, a car does not offer suggestions if self-driving is turned on – it drives itself. The outputs of the self-driving model are thus directly relevant to the driver's payoff, instead of being information. Furthermore, the purpose of self-driving is that the driver chooses a fixed action (whether to turn it on) instead of making decisions road condition by road condition.

[10]Recently, Liang, Lu and Mu (2023) elegantly examine algorithmic fairness in an information design setting, where a sender chooses inputs to an algorithm and a receiver chooses the algorithm.

experiment design and external validity," in *Handbook of Economic Field Experiments*, 1, 141–174: Elsevier.

Bansal, G., T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld (2021) "Does the Whole Exceed Its Parts? The Effect of AI Explanations On Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bohren, J. A. and D. N. Hauser (2024) "Misspecified Models in Learning and Games," *Annual Review of Economics*, 17.

Chassang, S. and S. Kapon (2022) "Designing Randomized Controlled Trials with External Validity in Mind," December, Working paper.

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012) "Selective trials: A principal-agent approach to randomized controlled experiments," *American Economic Review*, 102 (4), 1279–1309.

De-Arteaga, M., R. Fogliato, and A. Chouldechova (2020) "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Esponda, I. and D. Pouzo (2016) "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84 (3), 1093–1130.

Fudenberg, D., W. Gao, and A. Liang (2024) "How flexible is that functional form? Quantifying the restrictiveness of theories," *Journal of Economics and Statistics, Forthcoming*, Forthcoming.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022) "Measuring the Completeness of Economic Models," *Journal of Political Economy*, 130 (4), 956–990.

Fudenberg, D., G. Lanzani, and P. Strack (2021) "Limit points of endogenous misspecified learning," *Econometrica*, 89 (3), 1065–1098.

Fudenberg, D. and A. Liang (2020) "Machine Learning for Evaluating and Improving Theories," *ACM SIGecom Exchanges*, 18 (1), 4–11.

Gilboa, I. and D. Schmeidler (1989) "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18 (2), 141–153.

Iakovlev, A. and A. Liang (2023) "The Value of Context: Human versus Black Box Evaluators," December, Working paper.

Jussupow, E., I. Benbasat, and A. Heinzl (2020) "Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion," Working paper.

Kantorovich, L. V. and G. P. Akilov (1964) *Functional Analysis in Normed Spaces*: Pergamon Press.

Kasy, M. et al. (2013) "Why experimenters should not randomize, and what they should do instead," *European Economic Association & Econometric Society*, 1–40.

Lai, V. and C. Tan (2019) "On Human Predictions With Explanations and Predictions of Machine Learning Models: A Case Study On Deception Detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Liang, A., J. Lu, and X. Mu (2023) "Algorithm Design: A Fairness-Accuracy Frontier," Working paper.

Ludwig, J. and S. Mullainathan (2021) "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System," *Journal of Economic Perspectives*, 35 (4), 71–96.

Lundberg, S. M. and S.-I. Lee (2017) "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, 30.

Montiel Olea, J. L., P. Ortoleva, M. M. Pai, and A. Prat (2022) "Competing Models," *The Quarterly Journal of Economics*, 137 (4), 2419–2457.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016) ""Why Should I Trust You?" Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Savage, L. J. (1951) "The theory of statistical decision," *Journal of the American Statistical association*, 46 (253), 55–67.

Schwartzstein, J. and A. Sunderam (2021) "Using Models to Persuade," *American Economic Review*, 111 (1), 276–323.

Wald, A. (1949) "Statistical decision functions," *The Annals of Mathematical Statistics*, 165–205.

Proofs

Proof of Theorem 1 Suppose that $\overline{\Gamma}(f) = \overline{\Gamma}(\hat{f})$. Since $\overline{\Gamma}$ is the orthogonal projection onto Φ, and since the constant function $\mathbf{1}$ is contained in Φ,

$$\langle \mathbf{1}, f - \overline{\Gamma}(f) \rangle = \langle \mathbf{1}, \hat{f} - \overline{\Gamma}(\hat{f}) \rangle = 0\,.$$

Hence,

$$\mathbb{E}_{x\sim\mu_0}[\hat{f}(x)] = \langle \mathbf{1}, \hat{f} \rangle = \langle \mathbf{1}, \overline{\Gamma}(\hat{f}) \rangle = \langle \mathbf{1}, \overline{\Gamma}(f) \rangle = \langle \mathbf{1}, f \rangle = \mathbb{E}_{x\sim\mu_0}[f(x)]\,,$$

and thus

$$\begin{aligned} U(\hat{f}, a) = \mathbb{E}_{x\sim\mu_0}[u(x, \hat{f}(x), a)] =& w_0(a) + \mathbb{E}_{x\sim\mu_0}[x] \cdot w_1(a) + \mathbb{E}_{x\sim\mu_0}[\hat{f}(x)] \cdot w_2(a) \\ =& w_0(a) + \mathbb{E}_{x\sim\mu_0}[x] \cdot w_1(a) + \mathbb{E}_{x\sim\mu_0}[f(x)] \cdot w_2(a) \\ =& \mathbb{E}_{x\sim\mu_0}[u(x, f(x), a)] = U(f, a)\,, \end{aligned}$$

as desired. ■

Proof of Theorem 2 (OLS Is All You Need) Suppose $\phi \in \Phi$ and $f \in \overline{\Gamma}^{-1}(\phi)$. Since $\overline{\Gamma}$ is idempotent, $\overline{\Gamma}(\phi) = \phi = \overline{\Gamma}(f)$. Then by Theorem 1, for each $a \in A$, $U(\phi, a) = U(f, a)$. Then

$$U(f, a^*(\phi)) = U(\phi, a^*(\phi)) = \max_{a\in A} U(\phi, a) = \max_{a\in A} U(f, a) = \overline{U}(a),$$

as desired. ■

Lemma 1. *Let $\{g_i : X \to Y\}_{i=1}^N$ be linearly independent. There exist $\{x_i\}_{i=1}^N \subset X$ such that the matrix $G_N(\{x_i\}_{i=1}^N) = \begin{bmatrix} g_1(x_1) & \cdots & g_N(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_N) & \cdots & g_N(x_N) \end{bmatrix}$ has full rank.*

Proof. We proceed by induction on N.

Initial step: $N = 1$. By assumption, $g_1 \neq 0$. Then there exists x_1 such that $g_1(x_1) \neq 0$, as desired.

Induction step: $N > 1$. Suppose that the statement holds for $N-1$, and let $\{x_i\}_{i=1}^{N-1}$ be such that $G_{N-1}(\{x_i\}_{i=1}^{N-1})$ has full rank. Suppose toward a contradiction that for all $x_N \in X$, $G_N(\{x_i\}_{i=1}^N)$ does not have full rank. Then for any $x_N \in X$, there exist $\{z_i(x_N)\}_{i=1}^{N-1}$ such that

$$\begin{bmatrix} g_N(x_1) \\ \vdots \\ g_N(x_N) \end{bmatrix} = \sum_{i=1}^{N-1} z_i(x_N) \begin{bmatrix} g_i(x_1) \\ \vdots \\ g_i(x_N) \end{bmatrix}.$$

Then since $G_{N-1}(\{x_i\}_{i=1}^{N-1})$ has full rank, we must have

$$\begin{bmatrix} z_1(x_N) \\ \vdots \\ z_{N-1}(x_N) \end{bmatrix} = [G_{N-1}(\{x_i\}_{i=1}^{N-1})]^{-1} \begin{bmatrix} g_N(x_1) \\ \vdots \\ g_N(x_{N-1}) \end{bmatrix} =: \begin{bmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_{N-1} \end{bmatrix}.$$

Then for all $x_N \in X$, $g_N(x_N) = \sum_{i=1}^{N-1} \hat{z}_i g_i(x_N)$. Then $\{g_i\}_{i=1}^N$ are linearly dependent, a contradiction. ■

Proposition 2. *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$, that $\Phi = \hat{\Phi}^M$ for a subspace $\hat{\Phi}$ of $L^2(\mu_0)$, and that $\text{supp}\mu_0 = X$. For any (weak-$*$) open set $\mathcal{M} \subseteq \Delta(X)$ containing μ_0, there exists $\mu \in \mathcal{M}$ such that for all $\phi^* \in \Phi$, $w \in Y \setminus \{0\}$, and $z \in \mathbb{R}$, there exists $f \in \overline{\Gamma}_\mu^{-1}(\phi^*)$ such that $w \cdot \mathbb{E}_{x\sim\mu_0}[f(x)] = z$.*

Proof. Let $\{\phi_j\}_{j=1}^{\dim(\hat{\Phi})}$ be an orthonormal (in $L^2(\mu_0)$) basis for $\hat{\Phi}$; without loss let $\phi_1 = \mathbf{1}_X$. By Lemma 1, we can choose $\{x_i\}_{i=1}^{\dim(\hat{\Phi})} \subset X$ such that

$$\Psi = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_{\dim(\hat{\Phi})}(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_{\dim(\hat{\Phi})}) & \cdots & \phi_{\dim(\hat{\Phi})}(x_{\dim(\hat{\Phi})}) \end{bmatrix}$$

has full rank. Then since $\mathcal{M}$ is open in the weak-$*$ topology, there exists $\lambda \in (0,1)$ such that $\mu := \lambda\mu_0 + (1-\lambda)\sum_{i=1}^{\dim(\hat{\Phi})} \frac{1}{\dim(\hat{\Phi})}\delta_{x_i} \in \mathcal{M}$.

Now fix $w \in Y \setminus \{0\}$, and define $g^w \in F$ by

$$g_m^w(x) = w_m \left(1 - \sum_{i=1}^{\dim(\hat{\Phi})} \Psi_{1i}^{-1} \frac{1}{\mu_0(\{x_i\}) + \frac{1-\lambda}{\lambda \dim(\hat{\Phi})}} \mathbf{1}_{\{x_i\}}\right) \text{ for each } m \in \{1, \ldots, M\}.$$

For each $j \in \{1, \ldots, \dim(\hat{\Phi})\}$, g_m is orthogonal to ϕ_j in $L^2(\mu)$: For $j = 1$, we have

$$\begin{aligned}
\int_X g_m \phi_1 d\mu = \int_X g_m d\mu &= w_m \left(\lambda - \sum_{i=1}^{\dim(\hat{\Phi})} \Psi_{1i}^{-1} \frac{1}{\mu_0(\{x_i\}) + \frac{1-\lambda}{\lambda \dim(\hat{\Phi})}} \left(\lambda \mu_0(\{x_i\}) + (1-\lambda)\frac{1}{\dim(\hat{\Phi})}\right)\right) \\
&= w_m \lambda \left(1 - \Psi_{1\cdot}^{-1} \mathbf{1}\right) \\
&= w_m \lambda \left(1 - \Psi_{1\cdot}^{-1} \begin{bmatrix} \phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_{\dim(\hat{\Phi})}) \end{bmatrix}^\intercal\right) \\
&= w_m \lambda \left(1 - \Psi_{1\cdot}^{-1} \Psi_{\cdot 1}\right) = w_m \lambda (1 - 1) = 0.
\end{aligned}$$

For $j \neq 1$, since $\{\phi_j\}_{j=1}^{\dim(\hat{\Phi})}$ are orthonormal in $L^2(\mu_0)$, $\int_X \phi_j d\mu_0 = \int_X \phi_j \phi_1 d\mu_0 = 0$. Then we have

$$\begin{aligned}
\int_X g_m \phi_j d\mu &= w_m \left(\lambda \int_X \phi_j d\mu_0 - \sum_{i=1}^{\dim(\hat{\Phi})} \Psi_{1i}^{-1} \frac{1}{\mu_0(\{x_i\}) + \frac{1-\lambda}{\lambda \dim(\hat{\Phi})}} \left(\lambda \mu_0(\{x_i\}) + (1-\lambda)\frac{1}{\dim(\hat{\Phi})}\right) \phi_j(x_i)\right) \\
&= -w_m \lambda (\Psi_{1\cdot}^{-1} \Psi_{\cdot j}) = 0.
\end{aligned}$$

Then since $\{\phi_j\}_{j=1}^{\dim(\hat{\Phi})}$ is a basis for $\hat{\Phi}$, for each $m \in \{1, \ldots, M\}$, g_m is orthogonal to ψ in $L^2(\mu)$ for each $\psi \in \hat{\Phi}$. Then since $\Phi = \hat{\Phi}^M$, g is orthogonal to ϕ in $L^2(\mu)^M$ for each $\phi \in \Phi$. Consequently, we must have $\overline{\Gamma}_\mu(g) = 0$.

Moreover, $w \cdot \mathbb{E}_{x \sim \mu_0}[g] \neq 0$: We have

$$\begin{aligned}
w \cdot \int_X g d\mu_0 &= w \cdot w \left(1 - \sum_{i=1}^{\dim \Phi} \Psi_{1i}^{-1} \frac{1}{\mu_0(\{x_i\}) + \frac{1-\lambda}{\lambda \dim(\Phi)}} \mu_0(\{x_i\})\right) \\
&= ||w||^2 \left(1 - \Psi_{1\cdot}^{-1} \mathbf{1} + \frac{1-\lambda}{\lambda \dim(\Phi)} \Psi_{1\cdot}^{-1} \mathbf{1}\right) \\
&= ||w||^2 \left(1 - \Psi_{1\cdot}^{-1} \Psi_{\cdot 1} + \frac{1-\lambda}{\lambda \dim(\Phi)} \Psi_{1\cdot}^{-1} \Psi_{\cdot 1}\right) = ||w||^2 \left(1 - 1 + \frac{1-\lambda}{\lambda \dim(\Phi)}\right) > 0.
\end{aligned}$$

Now let $f = \phi^* + \frac{z - w \cdot \mathbb{E}_{x \sim \mu_0}[\phi^*(x)]}{w \cdot \mathbb{E}_{x \sim \mu_0}[g]} g$. Since $\overline{\Gamma}_\mu$ is linear and idempotent,

$$\overline{\Gamma}_\mu(f) = \overline{\Gamma}_\mu(\phi^*) + \frac{z - w \cdot \mathbb{E}_{x \sim \mu_0}[\phi^*(x)]}{w \cdot \mathbb{E}_{x \sim \mu_0}[g]} \overline{\Gamma}_\mu(g) = \overline{\Gamma}_\mu(\phi^*) = \phi^*.$$

So $f \in \overline{\Gamma}_\mu^{-1}(\phi^*)$. Moreover, we have

$$w \cdot \mathbb{E}_{x\sim\mu_0}[f(x)] = w \cdot \mathbb{E}_{x\sim\mu_0}[\phi^*(x)] + \frac{z - w \cdot \mathbb{E}_{x\sim\mu_0}[\phi^*(x)]}{w \cdot \mathbb{E}_{x\sim\mu_0}[g]} w \cdot \mathbb{E}_{x\sim\mu_0}[g] = z,$$

as desired. ■

Proof of Theorem 3 (Least-Squares Estimates Are Not Robust To Sampling Error) Let $\mu \in \mathcal{M}$ be the distribution whose existence is guaranteed by Proposition 2. By definition,

$$\inf_{f\in\overline{\Gamma}_\mu^{-1}} U(f,a) \geq \inf_{f\in F} U(f,a) \geq \inf_{y\in Y} \mathbb{E}_{x\sim\mu_0}[u(x,y,a)] \tag{4}$$

$$\Rightarrow \max_{a\in A} \inf_{f\in\overline{\Gamma}_\mu^{-1}} U(f,a) \geq \max_{a\in A} \inf_{f\in F} U(f,a) \geq \max_{a\in A} \inf_{y\in Y} \mathbb{E}_{x\sim\mu_0}[u(x,y,a)]. \tag{5}$$

Since payoffs are separable, we have $U(f,a) = w_0(a) + \mathbb{E}_{x\sim\mu_0}[x] \cdot w_1(a) + \mathbb{E}_{x\sim\mu_0}[f(x)] \cdot w_2(a)$. Then by Proposition 2, for any $\phi^* \in \Phi$, $\{U(f,a) | f \in \overline{\Gamma}_\mu^{-1}(\phi^*)\} = \mathbb{R}$, and so each of the quantities in (4) and (5) take the value $-\infty$. The claim follows.

Lemma 2. *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$, and hence $F = B_b(X)^M$. Let Γ be an explainer; let $\phi^* \in \Phi$ be a explanation; let $w \in Y \setminus \{0\}$ be a vector; let $z \in \mathbb{R}$. The set of priors*

$$\mathcal{M}_{z,w,\phi^*} := \{\mu \in \Delta(X) \mid \mathbb{E}_{x\sim\mu}[f(x)] \cdot w \neq z \forall f \in \Gamma^{-1}(\phi^*)\} \tag{6}$$

has finite dimension no greater than $\dim(\Phi)$.

Proof. Since $B_b(X)$ is complete in the sup-norm, so is $B_b(X)^M$ with the norm $||f|| = \sup_{x\in X} ||f(x)||$. For each $\mu \in \Delta(X)$, define the linear functional $e_{\mu,w}$ by $e_{\mu,w}(f) = \mathbb{E}_{x\sim\mu}[f(x)] \cdot w$; $e_{\mu,w}$ is continuous, since $|\mathbb{E}_{x\sim\mu}[f(x)] \cdot w| \leq \sup_{x\in X} |f(x) \cdot w| \leq ||w|| \cdot ||f||$. Choose a basis $\mathcal{B}$ of $\Gamma(F) = \Phi$. Suppose toward a contradiction that there is a finite linearly independent set $\mathcal{M} \subseteq \mathcal{M}_{y,w,\phi^*}$ with $|\mathcal{M}| > \dim(\Phi)$.[11] We first prove three claims.

Claim L2.1: $\Phi + \ker(\Gamma) = F$**.** By definition, $\Phi + \ker(\Gamma) \subseteq F$. Since $\mathcal{B}$ is a basis for Φ, for every $f \in F$, there exist $\{c_\phi\}_{\phi\in\mathcal{B}} \subset \mathbb{R}$ such that $\Gamma(f) = \sum_{\phi\in\mathcal{B}} c_\phi \phi$. Then since Γ is idempotent, $\Gamma(\sum_{\phi\in\mathcal{B}} c_\phi\phi) = \sum_{\phi\in\mathcal{B}} c_\phi \Gamma(\phi) = \sum_{\phi\in\mathcal{B}} c_\phi\phi = \Gamma(f)$. Then $f - \sum_{\phi\in\mathcal{B}} c_\phi\phi \in \ker(\Gamma)$, and hence $f \in \Phi + \ker(\Gamma)$.

Claim L2.2: For any $\mu \in \mathcal{M}_{z,w,\phi^*}$**,** $\ker(\Gamma) \subseteq \ker(e_{\mu,w})$**:** Suppose not. Then there exists $g \in \ker(\Gamma)$ such that $\mathbb{E}_{x\sim\mu}[g(x)] \cdot w \neq 0$. Then for any $h \in \Gamma^{-1}(\phi^*)$, $f = h + \frac{(z - \mathbb{E}_{x\sim\mu}[h(x)]\cdot w)}{\mathbb{E}_{x\sim\mu}[g(x)]\cdot w} g \in \Gamma^{-1}(\phi^*)$. Then $\mathbb{E}_{x\sim\mu}[f(x)] \cdot w = z$, a contradiction.

[11] Assuming that $\mathcal{M}$ is finite is without loss, since if $|\mathcal{M}| = \infty$, we can always take a finite subset.

Claim L2.3: For each each $\mu \in \mathcal{M}$, there exists $g^\mu \in F$ such that $\mathbb{E}_{x\sim\mu}[g^\mu(x)] \cdot w = 1$ but $\mathbb{E}_{x\sim\mu'}[g^\mu(x)] \cdot w = 0$ for each $\mu' \in \mathcal{M} \setminus \{\mu\}$. Let $e_{-\mu,w} = \bigoplus_{\mu' \in \mathcal{M}\setminus\{\mu\}} e_{\mu',w} \in \mathcal{B}(F, \mathbb{R}^{|\mathcal{M}|-1})$ be the direct sum of the expectation functionals $e_{\mu',w}$ for the priors in $\mathcal{M}$ other than μ. Let $e^*_{-\mu,w} : \mathbb{R}^{|\mathcal{M}|-1} \to F^* = \mathcal{B}(F, \mathbb{R})$ be the adjoint of $e_{-\mu,w}$ defined by $e^*_{-\mu,w}(z) = z \cdot e_{-\mu,w}$. Note that $\{g : X \to \mathbb{R} : g = w \cdot f \text{ for some } f \in F\} = B_b(X)$; since $\mathcal{M}$ is linearly independent, and $B_b(X)$ contains the set of simple functions, $\{e_{\mu',w}\}_{\mu' \in \mathcal{M}}$ must be linearly independent as well. Consequently, $e_{\mu,w} \notin e^*_{-\mu,w}(\mathbb{R}^{|\mathcal{M}|-1})$.

Since it is a subspace of the finite-dimensional space $\mathbb{R}^{|\mathcal{M}|-1}$, $e_{-\mu,w}(F)$ is closed. Then by Kantorovich and Akilov (1964) Theorem 3* (2.XII), $e^*_{-\mu,w}(\mathbb{R}^{|\mathcal{M}|-1}) = \perp \ker(e_{-\mu,w}) = \{A \in F^* | A(f) = 0 \forall f \in \ker(e_{-\mu,w})\}$. It follows that there exists $g \in \ker(e_{-\mu,w}) = \bigcap_{\mu' \in \mathcal{M}\setminus\{\mu\}} \ker(e_{\mu',w})$ such that $e_{\mu,w}(g) \neq 0$; the claim follows by letting $g^\mu = \frac{1}{\mathbb{E}_{x\sim\mu}[g(x)]\cdot w} g$.

We now construct a function for each $m \in \mathbb{R}^{\mathcal{M}}$ that is in Φ, and which returns the μth entry of m when $e_{w,\mu}$ is applied to it.

For any $m \in \mathbb{R}^{\mathcal{M}}$, let $f^m(x) = \sum_{\mu\in\mathcal{M}} m_\mu g^\mu(x)$, where g^μ are as defined in Claim L2.3. Since F is a vector space, we must have $f^m \in F$. Then by Claim L2.1, $f^m = \phi^m + h^m$ where $\phi^m \in \Phi$ and $h^m \in \ker(\Gamma)$, and hence by Claim L2.2, $h^m \in \ker(e_{\mu,w})$ for each $\mu \in \mathcal{M}$. Then for each $\mu \in \mathcal{M}$, we have $m_\mu = e_{\mu,w}(f^m) = e_{\mu,w}(\phi^m) + e_{\mu,w}(h^m) = e_{\mu,w}(\phi^m)$.

Now for each $\phi \in \mathcal{B}$, let $z^\phi \in \mathbb{R}^{\mathcal{M}}$ be the vector whose μth entry is $z^\phi_\mu = e_{\mu,w}(\phi)$.

Claim L2.4: For each $m \in \mathbb{R}^{\mathcal{M}}$, $m \in \text{span}\{z^\phi\}_{\phi\in\mathcal{B}}$. Given $m \in \mathbb{R}^{\mathcal{M}}$, we can write $\phi^m = \sum_{\phi\in\mathcal{B}} \lambda^m_\phi \phi$ for some $\{\lambda^m_\phi\}_{\phi\in\Phi} \subseteq \mathbb{R}$. Then for each $\mu \in \mathcal{M}$, $m_\mu = \mathbb{E}_{x\sim\mu}[\phi^m(x)] \cdot w = \sum_{\phi\in\mathcal{B}} \lambda^m_\phi \mathbb{E}_{x\sim\mu}[\phi(x)] \cdot w$, and hence $m_\mu = \sum_{\phi\in\mathcal{B}} \lambda^m_\phi z^\phi_\mu$. It follows that $m = \sum_{\phi\in\mathcal{B}} \lambda^m_\phi z^\phi$, and hence $m \in \text{span}\{z^\phi\}_{\phi\in\mathcal{B}}$.

We now complete the proof. Since $\mathcal{B}$ is a basis for Φ, it has no more than $\dim(\Phi)$ elements; it follows from Claim L2.4 that $\dim(\mathbb{R}^M) = |\mathcal{B}| \leq \dim(\Phi) < |\mathcal{M}|$, a contradiction. ■

Proof of Proposition 1 Follows immediately from Lemma 2 by identifying each $x \in X$ with the degenerate distribution δ_x with $\delta_x(\{x\}) = 1$. ■

Proof of Theorem 4 (Worst-Case Model Outcomes are Inexplicable) Fix $\phi \in \Gamma(F)$, and for each $w \in Y$ and $z \in \mathbb{R}$, let $X_{z,w,\phi} \equiv \{x \in X \mid f(x) \cdot w \neq z \forall f \in \Gamma^{-1}(\phi)\}$. By Proposition 1, each $X_{z,w,\phi}$ is finite. Since X is convex, it has no isolated points, so it follows that for each z, $X \setminus X_{z,w,\phi}$ is dense in X. Then since u is continuous, for each $w \in Y$, $z \in \mathbb{R}$, and $a \in A$, $u(X, y, a) \subseteq \text{cl}(u(X \setminus X_{z,w,\phi}, y, a))$. Hence $\inf_{x\in X} u(x, y, a) \geq \inf_{x\in X\setminus X_{z,w,\phi}} u(x, y, a)$, and since $X \setminus X_{z,w,\phi} \subseteq X$, we have

$$\inf_{x\in X} u(x, y, a) = \inf_{x\in X\setminus X_{z,w,\phi}} u(x, y, a).$$

Since u depends on one dimension of output for any given action, we have $u(x,y,a) = v(x, w(a)\cdot y, a)$. Then by definition of $X_{z,w,\phi}$, for each $y \in Y$ and $a \in A$,

$$\begin{aligned}\{u(x,y,a) \mid x \in X \setminus X_{w(a)\cdot y, w(a),\phi}\} &= \{v(x, y\cdot w(a), a) \mid f \in \Gamma^{-1}(\phi), x \in X \setminus X_{w(a)\cdot y, w(a),\phi}\} \\ &= \left\{ v(x, f(x)\cdot w(a), a) \;\middle|\; \begin{array}{c} f \in \Gamma^{-1}(\phi), \\ x \in X \setminus X_{z,w(a),\phi}, f(x)\cdot w(a) = y \cdot w(a) \end{array} \right\} \\ &\subseteq \{u(x, f(x), a) \mid f \in \Gamma^{-1}(\phi), x \in X\}.\end{aligned}$$

It follows that for each $y \in Y$ and $a \in A$,

$$\inf_{x\in X} u(x,y,a) = \inf_{x \in X \setminus X_{z,w(a),\phi}} u(x,y,a) \geq \inf_{\substack{x\in X \\ f\in\Gamma^{-1}(\phi)}} u(x, f(x), a).$$

Taking infima over y yields $\underline{R}(a) = \inf_{x\in X, y\in Y} u(x,y,a) \geq \inf_{x\in X, f\in\Gamma^{-1}(\phi)} u(x,f(x),a) = R(\phi, a|\Gamma)$. Then since $\Gamma^{-1}(\phi) \subseteq F$ and $\{u(x,f(x),a) \mid f \in F\} \subseteq \{u(x,y,a) \mid y \in Y\}$, we have

$$\inf_{\substack{x\in X \\ f\in\Gamma^{-1}(\phi)}} u(x,f(x),a) \geq \inf_{\substack{x\in X \\ f\in F}} u(x,f(x),a) \geq \inf_{\substack{x\in X \\ y\in Y}} u(x,y,a) \geq \inf_{\substack{x\in X \\ f\in\Gamma^{-1}(\phi)}} u(x,f(x),a),$$

and so all the quantities must be equal. Hence $\underline{R}(a) = R(\phi, a|\Gamma) = \inf_{\substack{x\in X \\ y\in Y}} u(x,y,a)$, as desired. (2) follows by taking maxima over A. ■

Proof of Theorem 5 (Futility of Explanation with Ambiguity Aversion) Fix $\phi \in \Gamma(F)$, and for each $z \in \mathbb{R}$ and $w \in Y$, let $\mathcal{M}_{z,w,\phi}$ be as in (6). By Lemma 2, for each $z \in \mathbb{R}$ and $w \in Y$, $\dim(\mathcal{M}_{z,w,\phi}) \leq \dim(\Phi) < \dim(\mathcal{M})$.

Claim T5.1. For each $z \in \mathbb{R}$ and $w \in Y$, $\mathcal{M} \setminus \mathcal{M}_{z,w,\phi}$ is dense in $\mathcal{M}$ (in the weak*-topology). Since $\dim(\mathrm{aff}(\mathcal{M}_{z,w,\phi})) = \dim(\mathcal{M}_{z,w,\phi}) < \dim(\mathcal{M})$, $\mathcal{M} \setminus \mathrm{aff}(\mathcal{M}_{z,w,\phi})$ is nonempty. Given $\mu \in \mathcal{M}_{z,w,\phi}$, choose $\mu' \in \mathcal{M} \setminus \mathrm{aff}(\mathcal{M}_{z,w,\phi})$. Then for each n, $\mu_n = \frac{1}{n}\mu' + (1-\frac{1}{n})\mu \in \mathcal{M}$ (since $\mathcal{M}$ is convex) but $\mu_n \notin \mathrm{aff}(\mathcal{M}_{z,w,\phi})$ (since if it was, then because $\mu \in \mathrm{aff}(\mathcal{M}_{z,w,\phi})$, we would have to have $\mu' = n\mu_n - (n-1)\mu \in \mathrm{aff}(\mathcal{M}_{z,w,\phi})$)). Since $\mu_n \to_{w^*} \mu$, μ is a limit point of $\mathcal{M} \setminus \mathcal{M}_{z,w,\phi}$; the claim follows.

Since u is continuous, for each $w, y \in Y$, $z \in \mathbb{R}$, and $a \in A$, it follows from Claim T5.1 that $\{\mathbb{E}_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M}\} \subseteq \mathrm{cl}(\{\mathbb{E}_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M} \setminus \mathcal{M}_{z,w,\phi}\})$. Hence $\inf_{\mu\in\mathcal{M}} \mathbb{E}_{x\sim\mu}[u(x,y,a)] \geq \inf_{\mu\in\mathcal{M}\setminus\mathcal{M}_{z,w,\phi}} \mathbb{E}_{x\sim\mu}[u(x,y,a)]$, and since $\mu \in \mathcal{M} \setminus \mathcal{M}_{z,w,\phi} \subseteq \mathcal{M}$, we have

$$\inf_{\mu\in\mathcal{M}} \mathbb{E}_{x\sim\mu}[u(x,y,a)] = \inf_{\mu\in\mathcal{M}\setminus M_{z,w,\phi}} \mathbb{E}_{x\sim\mu}[u(x,y,a)].$$

Since u is separable, we have $u(x,y,a) = w_0(a) + w_1(a)\cdot x + w_2(a)\cdot y$. Then by definition

of $\mathcal{M}_{z,w,\phi}$, for each $y \in Y$ and $a \in A$,

$$\begin{aligned}
&\{\mathbb{E}_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M} \setminus \mathcal{M}_{w_2(a)\cdot y, w_2(a),\phi}\} \\
&\qquad = \{w_0(a) + w_1(a)\cdot \mathbb{E}_{x\sim\mu}[x] + w_2(a)\cdot y \mid f \in \Gamma^{-1}(\phi), \mu \in \mathcal{M} \setminus \mathcal{M}_{w_2(a)\cdot y, w_2(a),\phi}\} \\
&\qquad = \left\{ w_0(a) + w_1(a)\cdot \mathbb{E}_{x\sim\mu}[x] + w_2(a)\cdot \mathbb{E}_{x\sim\mu}[f(x)] \;\middle|\; \begin{array}{c} \mu \in \mathcal{M} \setminus \mathcal{M}_{w_2(a)\cdot y, w_2(a),\phi}, \\ f \in \Gamma^{-1}(\phi), \mathbb{E}_{x\sim\mu}[f(x)] = y \end{array} \right\} \\
&\qquad \subseteq \{\mathbb{E}_{x\sim\mu}[u(x,f(x),a)] \mid f \in \Gamma^{-1}(\phi), \mu \in \mathcal{M}\}.
\end{aligned}$$

It follows that for each $y \in Y$ and $a \in A$,

$$\inf_{\mu\in\mathcal{M}} \mathbb{E}_{x\sim\mu}[u(x,y,a)] = \inf_{\mu\in\mathcal{M}\setminus\mathcal{M}_{w_2(a)\cdot y, w_2(a),\phi}} \mathbb{E}_{x\sim\mu}[u(x,y,a)] \geq \inf_{\substack{\mu\in\mathcal{M} \\ f\in\Gamma^{-1}(\phi)}} \mathbb{E}_{x\sim\mu}[u(x,f(x),a)].$$

Taking infima over y yields $\inf_{\mu\in\mathcal{M},y\in Y} \mathbb{E}_{x\sim\mu}[u(x,y,a)] \geq \inf_{\mu\in\mathcal{M},f\in\Gamma^{-1}(\phi)} \mathbb{E}_{x\sim\mu}[u(x,f(x),a)]$.

Moreover, we have

$$\begin{aligned}
\{\mathbb{E}_{x\sim\mu}[u(x,f(x),a)] \mid \mu \in \mathcal{M}, f \in F\} &= \{\mathbb{E}_{x\sim\mu}[w_0(a) + w_1(a)\cdot x + w_2(a)\cdot f(x)] \mid \mu \in \mathcal{M}, f \in F\} \\
&= \{w_0(a) + w_1(a)\cdot \mathbb{E}_{x\sim\mu}[x] + w_2(a)\cdot \mathbb{E}_{x\sim\mu}[f(x)] \mid \mu \in \mathcal{M}, f \in F\} \\
&\subseteq \{w_0(a) + w_1(a)\cdot \mathbb{E}_{x\sim\mu}[x] + w_2(a)\cdot y \mid \mu \in \mathcal{M}, y \in Y\} \\
&= \{\mathbb{E}_{x\sim\mu}[u(x,f(x),a)] \mid \mu \in \mathcal{M}, y \in Y\}
\end{aligned}$$

Then we have (since $\Gamma^{-1}(\phi) \subseteq F$)

$$\inf_{\substack{\mu\in\mathcal{M} \\ f\in\Gamma^{-1}(\phi)}} \mathbb{E}_{x\sim\mu}[u(x,f(x),a)] \geq \inf_{\substack{\mu\in\mathcal{M} \\ f\in F}} \mathbb{E}_{x\sim\mu}[u(x,f(x),a)] \geq \inf_{\substack{\mu\in\mathcal{M} \\ y\in Y}} \mathbb{E}_{x\sim\mu}[u(x,y,a)] \geq \inf_{\substack{\mu\in\mathcal{M} \\ f\in\Gamma^{-1}(\phi)}} \mathbb{E}_{x\sim\mu}[u(x,f(x),a)],$$

and so all the quantities must be equal. Then taking maxima over A yields $R_{\mathcal{M}}(\phi|\Gamma) = \underline{R}_{\mathcal{M}} = \max_{a\in A} \inf_{\substack{\mu\in\mathcal{M} \\ y\in Y}} \mathbb{E}_{x\sim\mu}\left[u(x,y,a)\right]$, as desired. ■