Explaining Models*

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

November 29, 2024

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, ordinary 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. To focus on intelligibility as the main factor obscuring the model from the DM, we abstract away from any sampling error that might be involved in this process of explanation.

We require the mapping between the space of true models and the space of intelligible models—what we call an *explainer*—to obey two criteria. First, if the true model is already intelligible, the explainer should not explain it with a different model. Second, if the true model is a mixture of two models generated by a randomization device that is independent of the state (e.g., one model holding half the time; another model, the other half), then the true model's explanation should be a mixture of those two models' explanations. 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 linear regression in policy evaluation and local approximations in machine learning.

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, regu-

lators 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, we show that, for any true model, as the set of intelligible models becomes richer (but still finite-dimensional), it is possible to make the DM as well off as if he understood the true model itself, simply by explaining that model with the ordinary least squares (OLS) method (Theorem 2).[1]

Unlike when the DM cares about the average, we show that if the DM cares about their worst-case payoff, *any* explanation is no better than having *no explanation at all* (Theorem 3). 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 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 3 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 4 shows that any explainer is unhelpful.

Related Literature Using models to rationalize data has emerged as a central theme in recent microeconomic theory. Montiel Olea, Ortoleva, Pai and Prat (2022) demonstrate that simple models appear more predictive with small samples, while complex models prove better with large samples, while Spiegler (2016) and Schwartzstein and Sunderam (2021) explore how people use models to interpret data and persuade others. This paper instead focuses on explaining complex true models with simpler ones for decision-making purposes.[2]

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

[2] Blattner, Nelson and Spiess (2021) examine similar trade-offs in a principal-agent context.

Research by Fudenberg and Liang (2020); Fudenberg, Kleinberg, Liang and Mullainathan (2022); Andrews, Fudenberg, Liang and Wu (2023); Fudenberg, Gao and Liang (2024) investigates machine learning's role in enhancing behavioral models and economic theories. We elaborate on the connections between this paper and that work in the model section.

The presence of potentially incorrect models in our framework naturally links to studies on misspecified models (Hansen and Sargent 2001; Esponda and Pouzo 2016), though unlike typical work in this area (Frick, Iijima and Ishii 2020; Fudenberg, Lanzani and Strack 2021; He and Libgober 2024), our DM neither updates beliefs nor learns dynamically. Instead, we examine whether a DM with a potentially incorrect model can achieve near-optimal utility.

This paper relates to statistical decision theory (Wald 1949; Savage 1951), though here the unknown object is the true model itself rather than a parameter. While traditional statistical decision theory focuses on direct data interpretation (Berger 2013), our framework requires understanding data through intelligible explanations.

The implications for policy evaluation situate our research within microeconomic theory involving policy choices and RCTs (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). Unlike Banerjee et al. (2020), we focus not on sampling error but on DMs' inability to fully comprehend true models even with infinite data, illuminating theoretical boundaries of common evaluation methods.

By examining models as decision aids, our analysis intersects with research on human reliance on AI models (Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan 2018; Athey, Bryan and Gans 2020; Agrawal, Gans and Goldfarb 2022), where studies show people resist using models they don't understand (Yeomans, Shah, Mullainathan and Kleinberg 2019; Chen, Feng, Sharma and Tan 2023). Our results identify when model explanations can prove useful or ineffective.

The challenge of model interpretability positions our work alongside research on "explainable AI" (Došilović, Brčić and Hlupić 2018; Lipton 2018; Molnar 2020), including methods like LIME (Local Interpretable Model-agnostic Explanations) by Ribeiro, Singh and Guestrin (2016). Our findings suggest OLS explanations can be nearly optimal for users who care about the average but unhelpful for those who care about the worst-case, aligning with research questioning the reliability of AI explanations (Rudin 2019; Lakkaraju and Bastani 2020; Slack, Hilgard, Jia, Singh and Lakkaraju 2020).

Outline The remainder of the paper proceeds as follows. Section 2 describes the paper's setting. Section 3 and Section 4 provide the main results, the former when the DM is Utilitarian and the latter when the DM is Rawlsian. Section 5 provides a discussion of the paper's findings. Section 6 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* and component $y \in Y$ is interpreted as an *output*. 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 most of our analysis, we assume that u is *separable*, in the sense that $u(x, y, a) = w_0(a)+x^\intercal w_1(a)+y^\intercal w_2(a)$ 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 function $f : X \to Y$. Given an input value $x \in X$, a true model f specifies the relationship between inputs and outputs via $y = f(x)$.[3]

Let $F \subseteq Y^X$ be the set of possible true models. Note that a true model could be highly complex: f could be nonlinear, discontinuous, non-differentiable, non-measurable, 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

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

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

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

[3]While we assume the true model f is a deterministic function from inputs to outputs, extra randomness can readily be incorporated into our framework by letting f be a function of both the inputs x and an independent randomization device $\varepsilon \in [0, 1]$.

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 (separable) 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*, which is defined below:

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 $\Gamma(f) \in \Phi$, so that the DM is able to understand the true model through this *explanation* $\Gamma(f)$. Linearity and idempotency are equivalent to requiring explainers to satisfy the following desirable properties:

1. (Consistency): $\Gamma(\phi) = \phi$ for all $\phi \in \Phi$. That is, if the true model f is intelligible, then the explainer should explain it by the true model itself.
2. (Law of Iterated Expectations): $\Gamma(\lambda \cdot g + (1-\lambda) \cdot h) = \lambda \cdot \Gamma(g) + (1-\lambda) \cdot \Gamma(h)$ for all $\lambda \in [0, 1]$ and for all $g, h \in F$. That is, if a model $f \in F$ is generated, via the law of iterated expectations, by mixing two models $g, h \in F$ using a randomization device $\varepsilon \in \{0, 1\}$ that is independent of the state (x, y) with $\mathbb{P}[\varepsilon = 1] = \lambda$:

$$f(x) = \mathbb{E}[y \mid x] = \lambda \mathbb{E}[y \mid x, \varepsilon = 1] + (1-\lambda)\mathbb{E}[y \mid x, \varepsilon = 0] = \lambda g(x) + (1-\lambda)h(x)\,,$$

then the explanation of f should not be affected by the randomization device and must

also satisfy the law of iterated expectations.[4]

For any explanation $\phi \in \Phi$ of a true model given by an explainer Γ, the true model may not be precisely the intelligible explanation ϕ. The set of possible models consistent with the explanation ϕ is given by

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

An explanation $\phi \in \Phi$ allows the DM to rule out all models that are inconsistent with the explanation ϕ. However, the DM is not able to identify which model $\hat{f} \in \Gamma^{-1}(\phi)$ is the true model that led to the explanation ϕ.

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

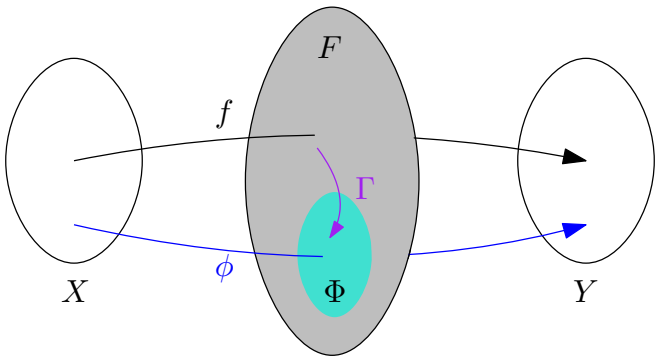

Figure 1: Models, explainers, and explanations. The figure depicts (1) the space F of possible true models f, which are functions from the space X of inputs to the space Y of outputs; (2) the subspace $\Phi \subset F$ of intelligible models ϕ; and (3) an explainer Γ that maps the space F of possible 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 economic theories (e.g., Fudenberg and Liang 2020; Fudenberg et al. 2022, 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.[5] Our

[4] In other words, the explainer Γ is not affected by extra randomization devices that are not part of the state space $X \times Y$.

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

key departure is that in their setting, prediction rules are *approximations*, and the goal is to approximate the truth as well as possible; in our setting, explanations are *information*, and the goal is to generate them in a way that is useful for 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 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 3) DM 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).

The assumption of separable payoffs ensures that for a decision maker who cares about their expected payoff — the case we consider in Theorems 1, 2, and 4 — the only payoff-relevant characteristic of the model is its expected output. In other words, the algebraic structure of the space of outputs Y that matters to the decision maker is exactly what's preserved by explainers satisfying the law of iterated expectations (property 2).

3 When are Explanations 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.[6] 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 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

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

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. The *ordinary least square (OLS)* explainer $\overline{\Gamma}$ is the unique orthogonal projection from F onto Φ. That is, for each $f \in F$, $\overline{\Gamma}(f)$ is the unique element of Φ such that $\langle \phi, f - \overline{\Gamma}(f) \rangle = 0$ for all $\phi \in \Phi$, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $L^2(\mu_0)^M$.[7]

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 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.

[7] Specifically, given a fixed prior μ_0 of input x, for any $f, g \in F$, define the inner product

$$\langle f, g \rangle := \mathbb{E}_{x \sim \mu_0} \left[\sum_{j=1}^{M} f_j(x) g_j(x) \right]\,.$$

As an immediate consequence of Theorem 1, it follows that under the OLS explainer $\overline{\Gamma}$, any utilitarian DM 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 $f \in F$, the DM's first-best value can be achieved by treating the explanation as the true model under the 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 the DM 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 if they understood the true model.

The results above suggest that explaining models can be very useful for a utilitarian decision maker who evaluates their payoffs using a fixed prior μ_0 over the inputs x. This, however, relies on the fact that the OLS explainer is constructed using the distribution μ_0, which is the same belief the DM uses to evaluate their payoff. In what follows, we show that the value of explaining models drastically decreases when the DM cares about the *worst case* rather than the *average* payoff.

4 When are Explanations Not Useful?

When the space of intelligible models is rich enough (but still finite dimensional), Theorem 2 shows that an ordinary least squares explanation achieves expected payoffs indistinguishable from those obtained with complete knowledge of the true model, for an expected utility maximizer with a fixed prior μ_0 over inputs. More precisely, 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 decision makers who care about the models' *worst-case*, rather than expected, performance. Policymakers may focus on those who could be disproportionately harmed by an intervention (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.

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 3 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 to a worst-case decision maker is futile. 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 $v : X \times Y \times A$.

Theorem 3 (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 useful information to a Rawlsian decision-maker: 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:

$$\max_{a \in A} R(\phi, a|\Gamma) = \max_{a \in A} \underline{R}(a) = \max_{a \in A} \inf_{\substack{y \in Y \\ x \in X}} u(x, y, a). \tag{1}$$

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., about the output that the true model produces for 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 $f(x) \cdot w = z$.*

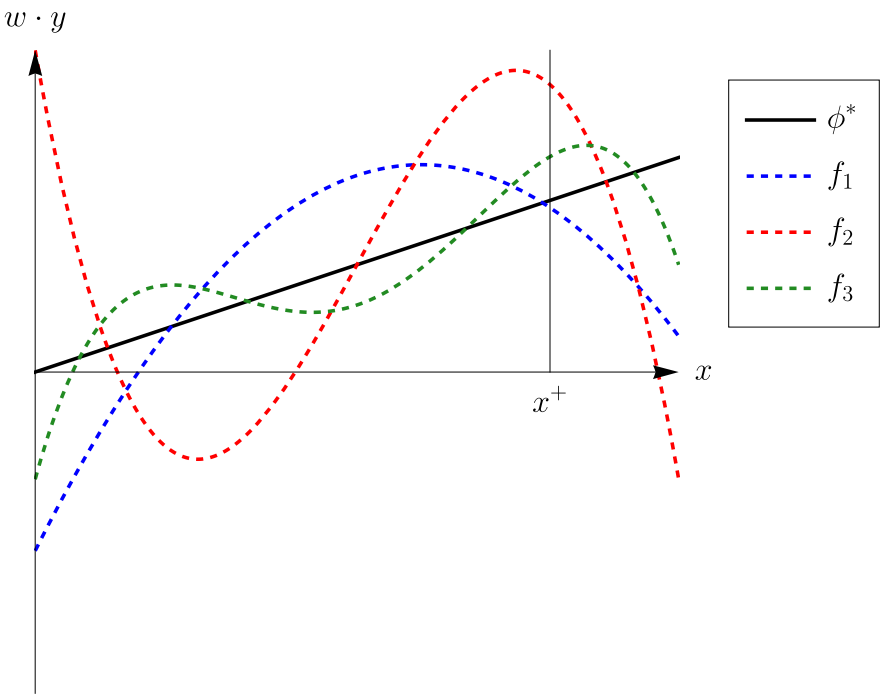

Figure 2: True models consistent with the same explanation. Proposition 1 shows that, at all but finitely many values of the input x, one can take any value z of a dimension of the output $w \cdot y$ and find a possible true model such that $w \cdot f(x) = z$ in a way consistent with the observed explanation. Here, we illustrate three distinct, possible true models that are each consistent with the explanation ϕ^*, under the OLS explainer and an input following a uniform distribution. OLS projects each possible true model to ϕ^*, but all three differ at all but finitely many input values. We highlight one input value, x^+, where the three possible true models differ.

Note that the impossibility result of Theorem 3 does not reverse as the dimension of Φ increases. In other words, no matter how many models are intelligible, *every* explainer is the same as no explanation for a Rawlsian DM. 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 Rawlsian DM, because the worst-case payoff consistent with an explanation is approximately the same, no matter the explanation.

Together, Theorem 3 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 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 3 shows, the finite-dimensional nature of the set of intelligible models makes this set unaffected by explanation.

To illustrate the implications of Theorem 3 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, if μ_0 represents the distribution of characteristics in a population, what if the DM does not know that distribution exactly?

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 3 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 4 shows that when the set of possible distributions is higher-dimensional than the space Φ of intelligible models, explanation is futile—even when, as in Section 3, the DM has separable payoffs, and thus cares only about the model's expected inputs and outputs.

Theorem 4 (Futility of Explanations 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 sufficiently rich set of priors, 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:

$$\max_{a \in A} R_{\mathcal{M}}(\phi, a|\Gamma) = \max_{a \in A} \underline{R}_{\mathcal{M}}(a) = \max_{a \in A} \inf_{\substack{y \in Y \\ \mu \in \mathcal{M}}} \mathbb{E}_{x \sim \mu}[u(x, y, a)]. \tag{2}$$

Theorem 4 shows that explanations of the true model are not helpful to an ambiguity averse DM. But it also provides a pessimistic perspective on Theorem 2. Even though explanation can work as well as understanding the true model for an expected utility maximizer, Theorem 4 shows that Theorem 2 relies on the DM's prior *exactly matching* the distribution of inputs μ_0 used to compute the model's OLS explanation. In particular, there are priors that are arbitrarily close to μ_0 for which the explanation is not almost perfect (as in Theorem 2), but instead useless.

5 Discussion

5.1 The Effectiveness of Explanations

Theorem 2 and Theorem 3 present a fundamental dichotomy between the two regimes when explaining complicated models. Explaining models using the canonical OLS approach is first-best when a decision maker cares about her average payoff, whereas (as long as the set of possible true models is rich enough) no explainer can improve the decisions of someone focused on the worst case.

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 3 suggests that when the policymaker is Rawlsian and cares about the worst-case outcome, it is impossible for any regression analyses to provide useful guidance for policymaking, 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 impossibility results (Theorems 3 and 4) 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 3 and 4 do not apply, and explanation may be possible even to a decision maker that cares about the worst case. 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 explanation to a decision maker that cares about the worst case is *never* possible, but that it is impossible *without the aid of theory*.

5.3 Recommendations vs. Explanations

Explanations fail to be useful in the contexts considered by Theorems 3 and 4 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 3 and 4: 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[8]

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

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 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).[9]

6 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

[8] Or in the ambiguity-averse case,

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

[9] 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.

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. We abstract from sampling error, but new insights might be gained by considering model explanation alongside model estimation. 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

Agrawal, A., J. Gans, and A. Goldfarb (2022) *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*: Harvard Business Press.

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.

Athey, S. C., K. A. Bryan, and J. S. Gans (2020) "The allocation of decision authority to human and artificial intelligence," in *AEA Papers and Proceedings*, 110, 80–84, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

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

Berger, J. O. (2013) *Statistical decision theory and Bayesian analysis*: Springer Science & Business Media.

Blattner, L., S. Nelson, and J. Spiess (2021) "Unpacking the black box: Regulating algorithmic decisions," Working paper.

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.

[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.

Chen, C., S. Feng, A. Sharma, and C. Tan (2023) "Machine explanations and human understanding," *Transactions on Machine Learning Research*, 1–30.

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.

Došilović, F. K., M. Brčić, and N. Hlupić (2018) "Explainable Artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 0210–0215, IEEE.

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

Frick, M., R. Iijima, and Y. Ishii (2020) "Misinterpreting others and the fragility of social learning," *Econometrica*, 88 (6), 2281–2328.

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.

Hansen, L. P. and T. J. Sargent (2001) "Robust control and model uncertainty," *American Economic Review*, 91 (2), 60–66.

He, K. and J. Libgober (2024) "Misspecified Learning and Evolutionary Stability,"Technical report, Working Paper.

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.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018) "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133 (1), 237–293.

Lakkaraju, H. and O. Bastani (2020) ""How do I Fool You?" Manipulating User Trust via

Misleading Black Box Explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.

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

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

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.

Molnar, C. (2020) *Interpretable machine learning*: Lulu. com.

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.

Rudin, C. (2019) "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, 1 (5), 206–215.

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.

Slack, D., S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju (2020) "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Spiegler, R. (2016) "Bayesian Networks and Boundedly Rational Expectations," *The Quarterly Journal of Economics*, 131 (3), 1243–1290.

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

Yeomans, M., A. Shah, S. Mullainathan, and J. Kleinberg (2019) "Making Sense of Recommendations," *Journal of Behavioral Decision Making*, 32 (4), 403–414.

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

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

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. *Suppose that F is the set of bounded Borel measurable functions from X to Y: $F = B_b(X)^M$. 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*

$$\mathcal{M}_{z,w,\phi^*} := \{\mu \in \Delta(X) \mid E_{x\sim\mu}[f(x)] \cdot w \neq z \forall f \in \Gamma^{-1}(\phi^*)\} \tag{3}$$

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 $e_{\mu,w}(f) = E_{x\sim\mu}[f(x)]\cdot w$; $e_{\mu,w}$ is continuous, since $|E_{x\sim\mu}[f(x)]\cdot w| \leq \sup_{x\in X} |f(x)\cdot w| \leq ||w||\cdot||f||$. 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 L1.1: $\Phi + \ker(\Gamma) = F$. 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 L1.2: 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 $E_{x\sim\mu}[g(x)]\cdot w \neq 0$. Then for any $h \in \Gamma^{-1}(\phi^*)$, $f = h + \frac{(z - E_{x\sim\mu}[h(x)]\cdot w)}{E_{x\sim\mu}[g(x)]\cdot w} g \in \Gamma^{-1}(\phi^*)$. Then $E_{x\sim\mu}[f(x)]\cdot w = z$, a contradiction.

Claim L1.3: For each each $\mu \in \mathcal{M}$, there exists $g^\mu \in F$ such that $E_{x\sim\mu}[g^\mu(x)]\cdot w = 1$ but $E_{x\sim\mu'}[g^\mu(x)]\cdot w = 0$ for each $\mu' \in \mathcal{M}\setminus\{\mu\}$. Let $e_{-\mu,w} = \bigoplus_{\mu'\in\mathcal{M}\setminus\{\mu\}} e_{\mu',w} \in \mathcal{B}(F, \mathbb{R}^{|\mathcal{M}|-1})$

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

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 $g^\mu = \frac{1}{E_{x \sim \mu}[g(x)] \cdot w} g$.

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 L1.3. Since F is a vector space, we must have $f^m \in F$. Then by Claim L1.1, $f^m = \phi^m + h^m$ where $\phi^m \in \Phi$ and $h^m \in \ker(\Gamma)$, and hence by Claim L1.2, $h^m \in \ker(e_{\mu,w})$ 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 L1.4: 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 $\mu \in \mathcal{M}$, $m_\mu = E_{x \sim \mu}[\phi^m(x)] \cdot w = \sum_{\phi \in \mathcal{B}} \lambda^m_\phi E_{x \sim \mu}[\phi(x)] \cdot w$, 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 L1.4 that $\dim(\mathbb{R}^M) = |\mathcal{B}| \leq \dim(\Phi) < |\mathcal{M}|$, a contradiction. ■

Corollary 1. *Suppose that F contains all bounded Borel measurable functions from $f : X \to Y$. 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*

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

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

Proof. Let $F' = B_b(X)^M \subseteq F$ and consider the restriction $\Gamma|_{F'}$. Then we have $\Gamma^{-1}(\phi^*) \supseteq \Gamma|^{-1}_{F'}(\phi^*)$. Hence

$$\mathcal{M}_{z,w,\phi^*} \equiv \{\mu \in \Delta(X) \mid E_{x \sim \mu}[f(x)] \cdot w \neq z \forall f \in \Gamma^{-1}(\phi^*)\} \subseteq \{\mu \in \Delta(X) \mid E_{x \sim \mu}[f(x)] \cdot w \neq z \forall f \in \Gamma|^{-1}_{F'}(\phi^*)\}.$$

The claim then follows immediately from Lemma 1. ■

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

Proof of Theorem 3 (Futility of Explanations in the Rawlsian Regime) 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 4 (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 (3). By Corollary 1, for each $z \in \mathbb{R}$ and $w \in Y$, $\dim(\mathcal{M}_{z,w,\phi}) \leq \dim(\Phi) < \dim(\mathcal{M})$.

Claim T4.1. 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 $\dim(\text{aff}(\mathcal{M}_{z,w,\phi})) = \dim(\mathcal{M}_{z,w,\phi}) < \dim(\mathcal{M})$, $\mathcal{M} \setminus \text{aff}(\mathcal{M}_{z,w,\phi})$ is nonempty.

Given $\mu \in \mathcal{M}_{z,w,\phi}$, choose $\mu' \in \mathcal{M} \setminus \text{aff}(\mathcal{M}_{z,w,\phi})$. Then for each n, $\mu_n = \frac{1}{n}\mu' + (1 - \frac{1}{n})\mu \in \mathcal{M}$ (since $\mathcal{M}$ is convex) but $\mu_n \notin \text{aff}(\mathcal{M}_{z,w,\phi})$ (since if it was, then because $\mu \in \text{aff}(\mathcal{M}_{z,w,\phi})$, we would have to have $\mu' = n\mu_n - (n-1)\mu \in \text{aff}(\mathcal{M}_{z,w,\phi})$)). 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 T4.1 that $\{E_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M}\} \subseteq \text{cl}(\{E_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M} \setminus \mathcal{M}_{z,w,\phi}\})$. Hence $\inf_{\mu\in\mathcal{M}} E_{x\sim\mu}[u(x,y,a)] \geq \inf_{\mu\in\mathcal{M}\setminus\mathcal{M}_{z,w,\phi}} E_{x\sim\mu}[u(x,y,a)]$, and since $\mu \in \mathcal{M} \setminus \mathcal{M}_{z,w,\phi} \subseteq \mathcal{M}$, we have

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

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 $a \in A$,

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

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

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

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

Moreover, we have

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

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

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

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