Explaining Models[*]

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

February 12, 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 finite-dimensional intelligible model. If the DM maximizes their average payoff, then an explanation using ordinary least squares is nearly 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.

JEL classification: C50, D81

Keywords: models, decision-making, model explanations, ordinary least squares

[*] We thank Mira Frick, Ryota Iijima, and Paul Goldsmith-Pinkham 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

[§] Yale School of Management, Email: alexander.zentefis@yale.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. Because several true models may be indistinguishable given an explanation, the DM evaluates their payoff according to the *worst-case* model that is consistent with the explanation when making a choice.

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, 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. We call this the *Utilitarian* regime. 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. We call this the *Rawlsian* regime.

The main results of the paper show that these two regimes have sharply contrasting implications for the usefulness of model explanations as decision aids. If the DM is Utilitarian, 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 arbitrarily close to as well off as understanding the true model itself, simply by explaining that model with the ordinary least squares (OLS) method (Theorem 1).[1] This result is not simply a consequence of the Stone-Weierstrass theorem. Theorem 1 not only relies on the convergence of the OLS explanation to a fixed true model, but also on the convergence of the *set* of true models consistent with a fixed *explanation*.

Unlike the Utilitarian regime, we show that if the DM is Rawlsian, *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.

The paper's two main results illustrate a fundamental dichotomy when it comes to model

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

explanations: an OLS-based explanation is nearly as good as knowing the true model if the DM cares about the average outcome, but no explanation is useful if the DM cares about the worst-case outcome.

Related Literature Several papers study models as devices that rationalize observed data. Montiel Olea, Ortoleva, Pai and Prat (2022) show theoretically that low-dimensional models can be believed to have superior predictive power when sample sizes are low, but high-dimensional models become accepted when sample sizes get large. Spiegler (2016) examines the implications of people having subjective causal models of long-run data distributions. Schwartzstein and Sunderam (2021) study situations where people use models to persuade others how to interpret data. In a series of papers, Fudenberg and Liang (2020); Fudenberg, Kleinberg, Liang and Mullainathan (2022); Andrews, Fudenberg, Liang and Wu (2023); Fudenberg, Gao and Liang (2024) evaluate how machine learning can enhance models of human behavior and improve economic theories. By contrast, this paper focuses on explaining unintelligible true models with intelligible models to facilitate decision-making.[2]

The paper's application to evaluating social policies ties the work to the branch of microeconomic theory that integrates aspects of decision theory with policy choices, particularly ones that rely on randomized controlled trials (RCTs) (Banerjee, Chassang and Snowberg 2017). See also Chassang, Padró i Miquel and Snowberg (2012), Kasy et al. (2013), Banerjee, Chassang, Montero and Snowberg (2020), and Chassang and Kapon (2022). Unlike Banerjee et al. (2020), this paper abstracts from the important issue of sampling error when estimating policy effects. The friction that we study is not the finite nature of the sample generated by the true model (the DGP), but the inability of a decision maker to fully understand the true model even with infinite data. The paper's setting captures situations in which social program evaluators can run RCTs, but also cases where they only have observational data at their disposal and instead use OLS-based methods to identify treatment effects. The paper's results illuminate the theoretical boundaries of these methods at explaining true underlying models.

The paper's application to the use of black-box AI models as decision aids connects it to the growing economics literature studying human reliance on AI models in decision-making (Kleinberg, Lakkaraju, Leskovec, Ludwig and Mullainathan 2018; Athey, Bryan and Gans 2020; Agrawal, Gans and Goldfarb 2022). Researchers have documented that people are reluctant to rely on models that they do not understand, even if those models prove highly accurate (Yeomans, Shah, Mullainathan and Kleinberg 2019; Chen, Feng, Sharma and Tan 2023). The paper's results reveal key differences between settings in which explanations of these models' predictions are useful, and settings in which they are not.

[2] Blattner, Nelson and Spiess (2021) study the trade-offs between complex models and explanations for them, though in a principal-agent setting where they analyze the optimal regulation of algorithms.

Indeed, a large literature in computer science has developed methods to explain the predictions of deep neural networks to make them more transparent to users, a field of research called *explainable AI* (Došilović, Brčić and Hlupić 2018; Lipton 2018; Molnar 2020). A very popular explanation method known as LIME (Local Interpretable Model-agnostic Explanations) approximates models with a linear regression around a data point to reveal features of the data having the largest impact on the model's prediction at that point (Ribeiro, Singh and Guestrin 2016). The paper's results suggest that if the AI model user is Utilitarian, then there exists an intelligible approximation of the model that is nearly as good as having a complete understanding of why the neural network makes the predictions it does. On the other hand, these sorts of explanations are unhelpful to a Rawlsian user of the model. This latter result mirrors findings in computer science that warn of the unreliability of AI model 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.

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 relation between inputs and outputs via $y = f(x)$.[3] Let $F \subseteq X^Y$ be a linear subspace that describes 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.

Actions and Payoffs A decision maker (henceforth DM) chooses an action a from a finite set $A = \{a_1, \dots, a_{|A|}\}$. The DM's payoff depends on the state of the world and the action

[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 $\varepsilon \in [0, 1]$.

chosen. Let $u : X \times Y \times A :\to \mathbb{R}$ denote the DM's payoff function, and assume that $u(\cdot, \cdot, a)$ is continuous on $X \times Y$ for all $a \in A$.

Example 1 (Treatment Effects: Banerjee et al. 2020)**.** 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 log likelihood ratios of the *treatment effects*, i.e., the success rates of each treatment, so that

$$y_a = \log\left(\frac{\mathbb{P}[\text{success} \mid a]}{1 - \mathbb{P}[\text{success} \mid a]}\right)$$

when the treatment is a. The policymaker's payoff is

$$u(x, y, a) = \mathbb{P}[\text{success} \mid a] = \text{logistic}(y_a) := \frac{1}{1 + e^{-y_a}}\,.$$

The log-likelihood ratio of success rate y_a under treatment a depends on the covariates x through a true model $f = (f_a)_{a \in A}$, so that $f_a(x) \in \mathbb{R}$ describes the log-likelihood ratio of the success rate of treatment a when the covariate is x.

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 potential improved traffic efficiency and the possibility of accidents). The regulator's payoff is

$$u(x, y, a) = \hat{u}(x, a) y_a - c(a)\,,$$

where $\hat{u}(x, a)$ is a cost-benefit multiplier that depends on condition x and rule a 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 $y = f(x)$, which is determined by the autonomous vehicle's algorithms, so that $f_a(x)$ 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. 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*, 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. (Mixture Invariance): $\Gamma(\lambda \cdot g + (1-\lambda) \cdot h) = \lambda \cdot \Gamma(g) + (1-\lambda) \cdot \Gamma(h)$ for all $\lambda \in [0, 1]$ and for all $g, h \in F$. That is, if model $f \in F$ is generated by mixing two models $g, h \in F$ using a randomization device that is independent of the state (x, y), then the explanation of f should not be affected by the randomization device and must also be the mixture of the explanations of g and h via the same randomization device.[4]

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

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

A class of explainers that will be of particular interest is the *ordinary least squares* explainers. We define these explainers in the context where F is a linear subspace of $L^2(\mu_0)^M$: the set of all measurable functions $f : X \to Y$ such that $\mathbb{E}[f_j(x)^2] < \infty$ for all $j \in \{1, \ldots, M\}$. For any $f, g \in F$, define the inner product $\langle f, g \rangle := \mathbb{E}[\sum_{j=1}^M f_j(x) g_j(x)]$.

Definition 2. Suppose that F is a linear subspace of $L^2(\mu_0)^M$. An explainer $\overline{\Gamma}$ for a decision problem (A, u, Φ) is the *ordinary least squares* explainer if $\overline{\Gamma}$ is orthogonal. That is, for any $\phi \in \Phi$, $\langle \phi, f - \overline{\Gamma}(f) \rangle = 0$.

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

Note that there is a unique ordinary least squares explainer of a decision problem (A, u, Φ), as orthogonal projection onto Φ is unique. The ordinary least square explainer $\overline{\Gamma}$ maps any true model f to the explanation $\overline{\Gamma}(f)$ that has the smallest distance to f (under the $L^2(\mu_0)$ norm) among all intelligible models $\phi \in \Phi$.

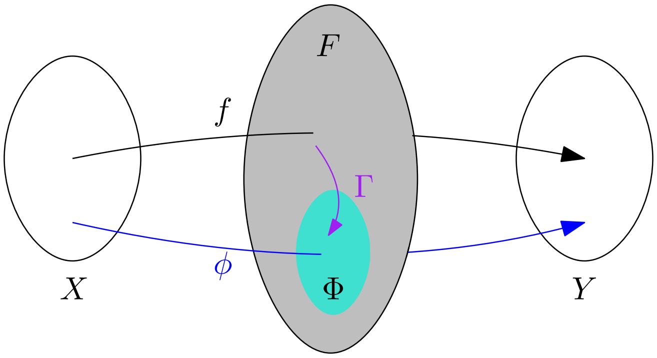

Figure 1: Illustration of the setting. Figure 1 shows (1) the space F of possible true models f, which are functions from the space X of inputs to the space Y of outputs; (2) the subspace $\Phi \subset F$ of intelligible models ϕ; and (3) an explainer Γ that maps the space F of possible true models to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

Utilitarian and Rawlsian Regimes We assume the DM is aware that an explanation ϕ may not be the same as the true model f, and thus evaluates their payoff based on the *worst-case* model that is consistent with a given explanation. We consider two regimes for how the DM uses the input distribution to evaluate their payoff: the *Utilitarian* regime and the *Rawlsian* regime.

Under the *Utilitarian* regime, the DM evaluates their payoff using the distribution μ_0 of inputs x. For any decision problem (A, u, Φ) and for any explainer Γ, the DM's payoff given an explanation $\phi \in \Phi$ is

$$U(\phi|\Gamma) := \max_{a \in A} \inf_{f \in \Gamma^{-1}(\phi)} \mathbb{E}[u(x, f(x), a)] .$$

In contrast, under the *Rawlsian* regime, the DM evaluates their payoff based on the *worst-case* input. For any decision problem (A, u, Φ) and for any explainer Γ, the DM's payoff given an explanation $\phi \in \Phi$ is

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

3 Utilitarian Regime

In many environments, decision makers care only about the *average* performance of their actions. 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 this section, we explore how explaining models can help the DM make better decisions under this *Utilitarian* regime. As a technical condition, we assume throughout this section that the set of all possible true models F is a linear subspace of $L^2(\mu_0)^M$.

For any true model $f \in F$, for any action set A and for any payoff $u : X \times Y \times A \to \mathbb{R}$, let $\overline{U}(f)$ be the payoff achieved by a Utilitarian DM who understands the true model f:

$$\overline{U}(f) := \max_{a \in A} \mathbb{E}[u(x, f(x), a)] .$$

In the Utilitarian regime, $\overline{U}(f)$ is the highest payoff that the DM can achieve, given that the true model is f. As a result, the performance of an explainer Γ for a given decision problem (A, u, Φ) can be evaluated by how close the DM's value $U(\Gamma(f)|\Gamma)$ given the true model f and the explainer Γ can be made to the benchmark $\overline{U}(f)$. Our next result suggests that, as long as the space of intelligible models Φ is large enough (but still finite-dimensional), the DM's value can be made arbitrarily close to $\overline{U}(f)$ by the *ordinary least squares* explainer.

Theorem 1 (Almost-Perfect Explanations)**.** *Consider any action set A, any $u : X \times Y \times A \to \mathbb{R}$ that is Lipschitz in y for all $x \in X$ and $a \in A$, and any nested sequence $\{\Phi_n\}$ of intelligible models such that $\dim(\Phi_n) = n$ for all $n \in \mathbb{N}$. For each $n \in \mathbb{N}$, let $\overline{\Gamma}_n$ be the ordinary least squares explainer for the decision problem (A, u, Φ_n). Then, for any true model $f \in F$ and for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that*

$$U(\overline{\Gamma}_n(f)|\overline{\Gamma}_n) > \overline{U}(f) - \varepsilon .$$

for all $n > N$.

To prove Theorem 1, it is necessary to show that as the dimensionality of intelligible models increases, there exists some intelligible model that approximates the true model. Having some finite-dimensional intelligible model that approximates the true model, however, is not enough to prove Theorem 1. For the DM's value to be arbitrarily close to the benchmark $\overline{U}(f)$, the set of all possible models that are consistent with the explaining model has to be small enough as well, so that the worst-case payoff would not be far from the benchmark payoff. While standard approximation methods (e.g., the Stone-Weierestrass theorem) could be used to show that certain true models (e.g., continuous functions f on X) can be approximated by some finite-dimensional intelligible models (e.g., polynomials), these approaches do not guarantee that all models consistent with the approximating intelligible models are close to the true model. Nonetheless, as we show in the proof, the orthogonality of the ordinary least squares explainer $\overline{\Gamma}$, together with the fact that F is a separable Hilbert space, allows us to find a sequence of intelligible models that approximates the true model *and* ensures that the sets of consistent models along the sequence eventually collapse to the true model.

In essence, Theorem 1 suggests that under the Utilitarian regime, regardless of the DM's payoff structure, the fact that some models are too complicated to be intelligible does not impose an intrinsic limitation on the DM's ability to make decisions. Rather, explaining a complex and unintelligible model to a Utilitarian DM could be very beneficial, and the benefit increases as the set of intelligible models widens. In the limit, by selecting a proper explainer—which, according to Theorem 1, can always be obtained via the standard OLS procedure—the DM can perform almost as well as if the true model were intelligible, however complicated it might be.

To illustrate the implications of Theorem 1, we may consider the treatment effect setting given by Example 1. Suppose that only polynomials are intelligible to the policymaker, and that the true model f—the underlying DGP that determines the log-likelihood ratio of a successful treatment—is highly complex and thus unintelligible. In this case, Φ_n is the set of n-degree polynomials, and $\overline{\Gamma}_n$ is the ordinary least squares regression of outputs onto n-degree polynomials. According to Theorem 1, as long as the degree n of intelligible polynomials is high enough, the ordinary least squares regression allows the policymaker to make treatment decisions arbitrarily well. In other words, running the OLS regression using high-enough degree polynomials is approximately optimal.

In fact, for a certain class of decision problems, explaining models can help the DM achieve *exactly* the first best payoff $\overline{U}(f)$. To see this, consider the class of decision problems (u, A, Φ), where u is affine in y:

$$u(x, y, a) = \sum_{j=1}^{M} \hat{u}_j(x, a) y_j + \hat{u}_0(x, a)\,,$$

for some $\{\hat{u}(\cdot, a)\}_{a\in A} \subseteq \Phi \subseteq L^2(\mu_0)^M$. That is, the DM's payoff is affine in outputs y, and the weights $\{\hat{u}(\cdot, a)\}_{a\in A}$, as functions in F, are intelligible. We refer to these decision problems as *payoff-intelligible affine decision problems.*

Theorem 2 (Perfect Explanations)**.** *For any payoff-intelligible affine decision problem (A, u, Φ) and for any true model $f \in F$, the DM's first-best value can be achieved by the ordinary least squares explainer $\overline{\Gamma}$. That is,*

$$U(\overline{\Gamma}(f)|\overline{\Gamma}) = \overline{U}(f)\,.$$

Theorem 2 shows that, when the DM's payoff is affine and when the weights $\{\hat{u}(\cdot, a)\}_{a\in A}$

are intelligible, the DM can do more than just perform approximately as well as knowing the true model through the ordinary least squares explainer: they can achieve *exactly* their first-best value $\overline{U}(f)$. In this case, the inability to understand complicated models does not impose any limitations on the DM to make an optimal decision, as long as the payoff-relevant weights $\{\hat{u}(\cdot, a)\}_{a \in A}$ are part of the set of intelligible models. The requirement that these weights are intelligible is not unrealistic, as it is typically the case that a DM understands their own ex-post payoffs. For instance, in the self-driving car example in Example 2, this requirement means that the multiplier function $\hat{u}(\cdot, a)$ is intelligible to the regulator for each rule a.

A special case of payoff-intelligible affine decision problems is when the DM's payoff is affine in both the inputs and the outputs. Namely, the DM has *separable* preferences of the form

$$u(x, y, a) = w_0(a) + x^\intercal w_1(a) + y^\intercal w_2(a)\,, \tag{1}$$

for some functions $w_0 : A \to \mathbb{R}$, $w_1 : A \to \mathbb{R}^K$, $w_2 : A \to \mathbb{R}^M$. In this case, as long as the set of intelligible models Φ contains a constant function, the decision problem (A, u, Φ) is payoff-intelligible and affine.

Corollary 1. *For any decision problem (A, u, Φ) where u is separable and Φ contains a constant function, the DM's first-best value can be achieved by the ordinary least squares explainer $\overline{\Gamma}$. That is,*

$$U(\overline{\Gamma}(f)|\overline{\Gamma}) = \overline{U}(f)\,.$$

4 Rawlsian Regime

When the space of intelligible models is rich enough (but still finite dimensional), Theorem 1 shows that it is possible to explain the model to a Utilitarian DM *almost perfectly* using ordinary least squares. In particular, the DM's expected payoff after observing the explanation produced by OLS is arbitrarily close to the expected payoff $\overline{U}(f)$ the DM would have if they knew and understood the true model itself. Moreover, when the decision problem is affine and payoff-intelligible, the OLS explainer allows a Utilitarian DM to make a choice *perfectly*.

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. Unfortunately, Theorem 3 below, in stark contrast to Theorem 1 and Theorem 2, reveals that explaining the true model to such a decision maker is futile. In particular, when the space of possible true models is rich enough, *no* explainer can do better than no explanation at all.

Theorem 3 (Futility of Explanations in the Rawlsian Regime)**.** *Suppose that F contains all bounded Borel measurable functions $f : X \to Y$. No explainer can provide useful information to a Rawlsian decision-maker: For any decision problem (A, u, Φ), any explainer Γ, and any $\phi \in \Phi$,*

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

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., 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 output y, and almost every possible input x, there is some model f with $f(x) = y$ 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 a explanation; let $y \in Y$ be an output. For all but finitely many $x \in X$, there exists $f \in \Gamma^{-1}(\phi^*)$ such that $f(x) = y$.*

Together, Theorem 3 and Proposition 1 reveal that explanations of complicated models offer no assistance to a Rawlsian DM at all, no matter how rich the set of (finite-dimensional) intelligible models is, and no matter how the DM's payoff is structured. The stark contrast between Theorem 1 and Theorem 3 stems from how the potential errors between the true model and an intelligible model are evaluated. When the DM is Utilitarian, even if the set of possible true models that is consistent with an explanation is infinite-dimensional, the *average* difference between any of the models in this set and the true model becomes arbitrarily small as the dimensionality of intelligible models increases. On the contrary, when the DM is Rawlsian, the average difference is irrelevant. Rather, the difference in the *worst case* scenario determines the performance of an explainer, which as Theorem 3 shows, is inherently limited by the finite-dimensionality of the set of intelligible models.

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.

So far, we have assumed that the DM 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? Theorem 4 shows that when the set of possible distributions is higher-dimensional than the space Φ of intelligible models, explanation is futile. This is true even when 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 $f : X \to Y$. If an ambiguity averse DM has a sufficiently rich set of priors, useful explanation is impossible, even with separable payoffs* (1)*: For any decision problem* (A, u, Φ)*, any explainer* Γ*, any convex* $\mathcal{M} \subseteq \Delta(X)$ *with* $\dim(\mathcal{M}) > \dim(\Phi)$ *and any* $\phi \in \Phi$*,*

$$R_{\mathcal{M}}(\phi|\Gamma) := \max_{a\in A} \inf_{\substack{\mu\in\mathcal{M} \\ f\in\Gamma^{-1}(\phi)}} E_{x\sim\mu}\left[u(x, f(x), a)\right] = \underline{R}_{\mathcal{M}} := \max_{a\in A} \inf_{\substack{\mu\in\mathcal{M} \\ f\in F}} E_{x\sim\mu}\left[u(x, f(x), a)\right]$$

$$= \max_{a\in A} \inf_{\substack{\mu\in\mathcal{M} \\ y\in Y}} E_{x\sim\mu}\left[u(x, y, a)\right].$$

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 1. Even though explanation can work arbitrarily well in the Utilitarian regime, Theorem 4 shows that Theorem 1 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 1), but instead useless.

5 Discussion

5.1 The Effectiveness of Explanations

Theorem 1 and Theorem 3 present a fundamental dichotomy over the effectiveness of explaining complicated models. Explaining models using the canonical OLS approach is approximately optimal when a decision maker is Utilitarian; whereas no explainer can improve a decision maker's choice if they are Rawlsian.

In the context of policymaking, Theorem 1 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 outcome, it is impossible for any regression analyses to provide useful guidance for policymaking. 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—does not alleviate those concerns.

5.2 Recommendations vs. Explanations

Useful explanations fail in Theorem 3 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 under the Rawlsian regime: 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[5]

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

[5] Or in the ambiguity-averse case,

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

Clearly, a recommender always gives the DM the full-information payoff $\overline{R}(f)$. 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 simply would follow the recommended action.

However, a recommendation only is 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).[6]

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 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.[7]

References

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

[6] 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.

[7] 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.

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.

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.

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.

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. 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.

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.

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.

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 (Almost Perfect Explanations) Consider any nested sequence

$\{\Phi_n\}$ of linear subspaces with $\dim(\Phi_n) = n$ for all $n \in \mathbb{N}$. We claim that there exists an orthonormal basis $\{\phi_n\}$ of F such that $\{\phi_i\}_{i=1}^n$ is an orthonormal basis of Φ_n. Since $\dim(\Phi_1) = 1$, $\Phi_1 = \text{span}\{\phi_1\}$ for some $\phi_1 \in F$ with $\|\phi_1\| = 1$. For each $n \in \mathbb{N}$, given an orthonormal basis $\{\phi_i\}_{i=1}^n$ of Φ_n, since Φ_n is a linear subspace of Φ_{n+1} and $\dim(\Phi_n) = n$, $\dim(\Phi_{n+1}) = n + 1$, $\Phi_{n+1} = \Phi_n \oplus \text{span}\{\phi_{n+1}\}$ for some $\phi_{n+1} \in \Phi_{n+1}$ such that $\langle \phi_i, \phi_{n+1} \rangle = 0$ for all $i \in \{1, \dots, n\}$ and $\|\phi_{n+1}\| = 1$. Thus, $\{\phi_i\}_{i=1}^{n+1}$ is an orthnormal basis of Φ_{n+1}. By induction, there exists an orthonormal sequence $\{\phi_n\}$ such that $\Phi_n = \text{span}\{\phi_i\}_{i=1}^n$ for all $n \in \mathbb{N}$. Moreover, since $\dim(\cup_{n=1}^\infty \Phi_n) = \infty$ and since F is a separable Hilbert space, F is isomorphic to $\cup_{n=1}^\infty \Phi_n$ and hence $\text{span}\{\phi_n\} = F$. Therefore, $\{\phi_n\}$ is an orthonormal basis of F.

Since $u(x, y, a)$ is Lipschitz, for any $x \in X$, $y, y' \in Y$, and $a \in A$, there exists $K(x, a)$ such that $|u(x, y, a) - u(x, y', a)| \leq K(x, a)\|y - y'\|$, and hence,

$$|\max_{a \in A} u(x, y, a) - \max_{a \in A} u(x, y', a)| \leq \max_{a \in A} |u(x, y, a) - u(x, y', a)| \leq \max_{a \in A} K(x, a)\|y - y'\|.$$

Meanwhile, note that for any finite dimensional subspace Φ, for any explainer Γ with $\text{Im}(\Gamma) = \Phi$, and for any $\phi \in \Phi$, we have

$$\max_{a \in A} \inf_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[u(x, \hat{f}(x), a)] \leq \max_{a \in A} \mathbb{E}[u(x, \phi(x), a)] = \mathbb{E}[u(x, \phi(x), a^*(\phi))] \, ,$$

for any $a^*(\phi)$ that maximizes $\mathbb{E}[u(x, \phi(x), a)]$, Therefore,

$$\inf_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[u(x, \hat{f}(x), a^*(\phi))] \leq \max_{a \in A} \inf_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[u(x, \hat{f}(x), a)] \leq \mathbb{E}[u(x, \phi(x), a^*(\phi))] \, .$$

Moreover, for any $\hat{f} \in \Gamma^{-1}(\phi)$ and for any $a \in A$,

$$\begin{aligned} |\mathbb{E}[u(x, \phi(x), a)] - \inf_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[u(x, \hat{f}(x), a)]| &= \sup_{\hat{f} \in \Gamma^{-1}(\phi)} |\mathbb{E}[u(x, \phi(x), a) - u(x, \hat{f}(x), a)]| \\ &\leq \sup_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[|u(x, \phi(x), a) - u(x, \hat{f}(x), a)|] \\ &\leq \sup_{\hat{f} \in \Gamma^{-1}(\phi)} \mathbb{E}[K(x, a)\|\hat{f}(x) - \phi(x)\|]. \end{aligned}$$

Together, for any $\phi \in \Phi$ and for any explainer Γ,

$$\begin{aligned}
&(\max_{a\in A}\mathbb{E}[u(x,f(x),a)] - \max_{a\in A}\inf_{\hat{f}\in\Gamma^{-1}(\phi)}\mathbb{E}[u(x,\hat{f}(x),a)])^2\\
=&(\max_{a\in A}\mathbb{E}[u(x,f(x),a)] - \max_{a\in A}\mathbb{E}[u(x,\phi(x),a)] + \mathbb{E}[u(x,\phi(x),a)] - \max_{a\in A}\inf_{\hat{f}\in\Gamma^{-1}(\phi)}\mathbb{E}[u(x,\hat{f}(x),a)])^2\\
\le&(\max_{a\in A}\mathbb{E}[K(x,a)\|f(x)-\phi(x)\|] + \mathbb{E}[u(x,\phi(x),a)] - \inf_{\hat{f}\in\Gamma^{-1}(\phi)}\mathbb{E}[u,\hat{f},a^*(\phi)])^2\\
\le&(\max_{a\in A}\mathbb{E}[K(x,a)\|f(x)-\phi(x)\|] - \sup_{\hat{f}\in\Gamma^{-1}(\phi)}\mathbb{E}[K(x,a)\|\hat{f}(x)-\phi(x)\|])^2\\
\le&(\max_{a\in A}\mathbb{E}[K(x,a)\|f(x)-\phi(x)\|])^2 + (\sup_{\hat{f}\in\Gamma^{-1}(\phi)}\mathbb{E}[K(x,a)\|\hat{f}(x)-\phi(x)\|])^2\\
\le&\max_{a\in A}\mathbb{E}[K(x,a)^2]\cdot\|f-\phi\|^2 + \max_{a\in A}\mathbb{E}[K(x,a)^2]\cdot\sup_{\hat{f}\in\Gamma^{-1}(\phi)}\|\hat{f}-\phi\|^2\,,
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality.

For any $n \in \mathbb{N}$, since $\overline{\Gamma}_n$ is the orthogonal projection onto $\Phi_n = \text{span}\{\phi_i\}_{i=1}^n$, and since $\{\phi_n\}_{n=1}^\infty$ is an orthonormal basis of F, for any true model f, $\overline{\Gamma}_n(f) = \sum_{i=1}^n \langle\phi_i, f\rangle\phi_i$, and therefore $\hat{f} \in \overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))$ if and only if $\langle\phi_i, f\rangle = \langle\hat{f}, \phi_i\rangle$ for all $i \in \{1,\dots,n\}$. It then follows that

$$\overline{\Gamma}_{n+1}^{-1}(\overline{\Gamma}_{n+1}(f)) \subseteq \overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))\,,$$

for all $n \in \mathbb{N}$. Therefore,

$$\sup_{\hat{f}\in\overline{\Gamma}_{n+1}^{-1}(\overline{\Gamma}_{n+1}(f))}\sum_{i=n+2}^\infty \langle\phi_i,\hat{f}\rangle^2 \le \sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))}\sum_{i=n+1}^\infty\langle\phi_i,\hat{f}\rangle^2$$

and hence $\lim_{n\to\infty}\sup_{\hat{f}\in\overline{\Gamma}_n^{*-1}(\overline{\Gamma}_n(f))}\sum_{i=n+1}^\infty\langle\phi_i,\hat{f}\rangle^2 = \kappa \ge 0$ exists. We now claim that $\kappa = 0$. Suppose the contrary, that $\kappa > 0$. Then for any $n \in \mathbb{N}$, there exists $f_n \in \overline{\Gamma}_n^{-1}(f)$ such that $\sum_{i=n+1}^\infty\langle\phi_i, f_n\rangle^2 > {}^{\kappa}\!/_2 > 0$. Moreover, since $\{\phi_n\}_{n=1}^\infty$ is an orthonormal basis of F,

$$\bigcap_{n=1}^\infty \overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f)) = \{f\}\,,$$

together with the fact that F is a complete metric space, $\lim_{n\to\infty}\|f_n - f\| = 0$. Since $\sum_{i=1}^\infty\langle\phi_i, f\rangle^2$ converges, there exists $N \in \mathbb{N}$ such that $\sum_{i=n+1}^\infty\langle\phi_i, f\rangle^2 < {}^{\kappa}\!/_4$ for all $n > N$, Therefore, for any $< N < m < n$

$$\begin{aligned}\sum_{i=m+1}^{\infty} |\langle \phi_i, f_n - f\rangle| \|f_n + f\| \geq & \sum_{i=n+1}^{\infty} \langle |\phi_i, f_n - f\rangle| \cdot |\langle \phi_i, f_n + f\rangle| \\ \geq & \sum_{i=n+1}^{\infty} \langle \phi_i, f_n\rangle^2 - \sum_{i=n+1}^{\infty} \langle \phi_i, f\rangle^2 \\ > & \frac{\kappa}{2} - \sum_{i=n+1}^{\infty} \langle \phi_i, f\rangle^2 \\ > & \frac{\kappa}{4} \\ > & 0\,,\end{aligned}$$

where the first inequality follows from Bessel's inequality. This leads to a contradiction, as $\lim_{n\to\infty} \|f_n - f\| = 0$ implies that

$$\limsup_{n\to\infty} \sum_{i=m+1}^{\infty} |\langle \phi_i, f - f_n\rangle| \|f_n + f\| = 0\,.$$

Together, we have

$$\lim_{n\to\infty} \sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \sum_{i=n+1}^{\infty} \langle \phi_i, \hat{f}\rangle^2 = \kappa = 0\,.$$

Now consider any $f \in F$ and any $\varepsilon > 0$. Since $\sum_{i=1}^{\infty} \langle \phi_i, f\rangle \phi_i = f$, there exists $N_1 \in \mathbb{N}$ such that

$$\|\overline{\Gamma}_n(f) - f\|^2 = \left\| \sum_{i=1}^{n} \langle \phi_i, f\rangle \phi_i - f \right\|^2 =< \frac{\varepsilon^2}{2 \max_{a\in A} \mathbb{E}[K(x,a)^2]}\,.$$

Also, since $\lim_{n\to\infty} \sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \sum_{i=n+1}^{\infty} \langle \phi_i, \hat{f}\rangle^2 = 0$, there exists $N_2 \in \mathbb{N}$ such that

$$\sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \sum_{i=n+1}^{\infty} \langle \phi_i, \hat{f}\rangle^2 < \frac{\varepsilon^2}{2 \max_{a\in A} \mathbb{E}[K(x,a)^2]}\,,$$

for all $n > N_2$. Let $N := \max\{N_1, N_2\}$, then for any $n > N$,

$$\begin{aligned}
&(\overline{U}(f) - U(\overline{\Gamma}_n(f)|\overline{\Gamma}_n))^2 \\
=&(\max_{a\in A} \mathbb{E}[u(x, f(x), a)] - \max_{a\in A} \inf_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \mathbb{E}[u(x, \hat{f}(x), a)])^2 \\
\leq& \max_{a\in A} \mathbb{E}[K(x,a)^2] \cdot \|f - \overline{\Gamma}_n(f)\|^2 + \max_{a\in A} \mathbb{E}[K^2(x,a)] \cdot \sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \|\hat{f} - \overline{\Gamma}_n(f)\|^2 \\
=& \max_{a\in A} \mathbb{E}[K(x,a)^2] \cdot \|f - \overline{\Gamma}_n(f)\|^2 + \max_{a\in A} \mathbb{E}[K(x,a)^2] \cdot \sup_{\hat{f}\in\overline{\Gamma}_n^{-1}(\overline{\Gamma}_n(f))} \sum_{i=n+1}^{\infty} \langle \phi, \hat{f} \rangle^2 \\
<&\frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2} \\
<&\varepsilon^2 ,
\end{aligned}$$

and thus $|\overline{U}(f) - U(\overline{\Gamma}_n(f)|\overline{\Gamma}_n)| < \varepsilon$, as desired. ■

Proof of Theorem 2 (Perfect Explanations) Consider any true model f and let $\phi := \overline{\Gamma}(f)$. Since $\overline{\Gamma}$ is the orthogonal projection onto $\Phi \supseteq \text{span}\{\hat{u}(\cdot, a)\}_{a\in A}$, for any $a \in A$,

$$\langle \hat{u}(\cdot, a), \hat{f} - \phi \rangle = 0 .$$

,for any $\hat{f} \in \Gamma^{*-1}(\phi)$, and hence,

$$\mathbb{E}\left[\sum_{j=1}^{M} \hat{u}_j(x,a)\hat{f}_j(x)\right] = \langle \hat{u}(\cdot, a), \hat{f} \rangle = \langle \hat{u}(\cdot, a), \phi \rangle = \mathbb{E}\left[\sum_{j=1}^{M} \hat{u}_j(x,a)\phi_j(x)\right] ,$$

for all $a \in A$. Therefore,

$$\begin{aligned}
\inf_{\hat{f}\in\Gamma^{*-1}(\phi)} \mathbb{E}[u(x, \hat{f}(x), a)] =& \mathbb{E}\left[\sum_{j=1}^{M} \hat{u}_j(x,a)\phi_j(x)\right] + \mathbb{E}[u_0(x,a)] \\
=& \mathbb{E}\left[\sum_{j=1}^{M} \hat{u}_j(x,a)f_j(x)\right] + \mathbb{E}[u_0(x,a)] \\
=& \mathbb{E}[u(x, f(x), a)] ,
\end{aligned}$$

for all $a \in A$, which implies

$$U(\overline{\Gamma}(f)|\overline{\Gamma}) = \overline{U}(f) ,$$

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 $y \in Y$ be an output.*

The set of priors

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

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 $||f|| = \sup_{x\in X, 1\leq i\leq M} |f(x)_i|$.[8] For each $\mu \in \Delta(X)$ and $1 \leq i \leq M$, define the entrywise expectation linear functional $e_{\mu,i}$ by $e_{\mu,i}(f) = E_{x\sim\mu}[f(x)]_i$; each $e_{\mu,i}$ is continuous, since $|E_{x\sim\mu}[f(x)]_i| \leq \sup_{x\in X} |f(x)_i| \leq ||f||$. Choose a basis $\mathcal{B}$ of $\Gamma(F) = \Phi$, and for each $\phi \in \mathcal{B}$, choose $f_\phi \in \Gamma^{-1}(\phi)$. Let $\mathcal{F} = \{f_\phi\}_{\phi\in\mathcal{B}}$. Suppose toward a contradiction that there is a finite linearly independent set $\mathcal{M} \subseteq \mathcal{M}_{y,\phi^*}$ with $|\mathcal{M}| > \dim(\Phi)$.[9] We first prove three claims.

Claim L1.1: $\operatorname{span}\mathcal{F} + \ker(\Gamma) = F$. By definition, $\operatorname{span}\mathcal{F} + \ker(\Gamma) \subseteq F$. Since $\mathcal{B}$ is a basis for $\Gamma(F)$, for every $f \in F$, there exist $\{c_\phi\}_{\phi\in\mathcal{B}} \subset \mathbb{R}$ such that $\Gamma(f) = \sum_{\phi\in\mathcal{B}} c_\phi \phi = \sum_{\phi\in\mathcal{B}} c_\phi \Gamma(f_\phi)$. Then $\Gamma(\sum_{\phi\in\mathcal{B}} c_\phi f_\phi) = \sum_{\phi\in\mathcal{B}} c_\phi \Gamma(f_\phi) = \Gamma(f)$. Then $f - \sum_{\phi\in\mathcal{B}} c_\phi f_\phi \in \ker(\Gamma)$, and hence $f \in \operatorname{span}\mathcal{F} + \ker(\Gamma)$.

Claim L1.2: For any $\mu \in \mathcal{M}_{y,\phi^*}$, $\ker(\Gamma) \cap \bigcap_{j\neq i} \ker(e_{\mu,j}) \subseteq \ker(e_{\mu,i})$ for some i: Suppose not. Then for each i, there exists $g^i \in \ker(\Gamma)$ such that $E_{x\sim\mu}[g^i(x)]_i \neq 0$ but $E_{x\sim\mu}[g^i(x)]_j = 0$ for each $j \neq i$. Then for any $h \in \Gamma^{-1}(\phi^*)$, $f = h + \sum_{i=1}^M \frac{(y_i - h(x)_i)}{E_{x\sim\mu}[g^i(x)]_i} g^i \in \Gamma^{-1}(\phi^*)$. Then $E_{x\sim\mu}[f(x)] = y$, a contradiction.

Claim L1.3: For each $1 \leq i \leq M$ and each $\mu \in \mathcal{M}$, there exists $g^i_\mu \in F$ such that $E_{x\sim\mu}[g^i_\mu(x)]_i = 1$ but $E_{x\sim\mu'}[g^i_\mu(x)]_i = 0$ for each $\mu' \in \mathcal{M}\backslash\{\mu\}$. Let $e_{-\mu,i} = \bigoplus_{\mu'\in\mathcal{M}\backslash\{\mu\}} e_{\mu',i} \in \mathcal{B}(F, \mathbb{R}^{|\mathcal{M}|-1})$ be the direct sum of the ith-entry expectation functionals for the priors in $\mathcal{M}$ other than μ. Let $e^*_{-\mu,i} : \mathbb{R}^{|\mathcal{M}|-1} \to F^* = \mathcal{B}(F, \mathbb{R})$ be the adjoint of $e_{-\mu,i}$ defined by $e^*_{-\mu,i}(z) = z \cdot e_{-\mu,i}$. Since $\mathcal{M}$ is linearly independent, and $F_i = B_b(X)$ contains the set of simple functions, $\{e_{\mu',i}\}_{\mu'\in\mathcal{M}}$ must be linearly independent as well; consequently, $e_{\mu,i} \notin e^*_{-\mu,i}(\mathbb{R}^{|\mathcal{M}|-1})$.

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

We now construct a function for each $z \in \mathbb{R}^{\mathcal{M}}$ that is in the span of $\mathcal{F}$, and which returns the μth element of z when its expectation is taken under μ.

By Claim L1.2, there exist $\{i_\mu\}_{\mu\in\mathcal{M}}$ such that for each $\mu \in \mathcal{M}$, $\ker(\Gamma) \cap \bigcap_{j\neq i_\mu} \ker(e_{\mu,j}) \subseteq$

[8] Note that since the Euclidean and taxicab metrics induce the same topology on $\mathbb{R}^M$, this norm induces the same topology on $B_b(X)^M$ as $||f|| = \sup_{x\in X} ||f(x)||$.

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

$\ker(e_{\mu,i_\mu})$. For any $z \in \mathbb{R}^{\mathcal{M}}$, let

$$f_z(x) = \sum_{\mu\in\mathcal{M}} z_\mu([g_\mu^{i_\mu}(x)]_{i_\mu} \oplus 0_{-i_\mu}),$$

where g_μ^i are as defined in Claim L1.3. For each $\mu \in \mathcal{M}$, since $g_\mu^{i_\mu} \in F = B_b(X)^M$, we must have $[g_\mu^{i_\mu}(x)]_{i_\mu} \in B_b(X)$, and hence $f_z \in F$. Then for each $\mu \in \mathcal{M}$, $E_{x\sim\mu}[f_z(x)] = z_\mu(1_{i_\mu} \oplus 0_{-i_\mu})$, and so $f_z \in \bigcap_{j\neq i_\mu} \ker(e_{\mu,j})$.

Claim L1.4: $f_z \in \operatorname{span}\mathcal{F}$. Suppose not. Then by Claim L1.1, $f_z \in \ker(\Gamma)$. Then we have $f_z \in \ker(\Gamma) \cap \bigcap_{j\neq i_\mu} \ker(e_{\mu,j})$, and hence $f_z \in \ker(e_{\mu,i_\mu})$, for each $\mu \in \mathcal{M}$. Then for each $\mu \in \mathcal{M}$, $E_{x\sim\mu}[f_z(x)] = z_\mu(1_{i_\mu} \oplus 0_{-i_\mu}) = 0$, and hence $z = 0$. But then $f_z = 0 \in \operatorname{span}\mathcal{F}$, a contradiction.

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

Claim L1.5: For each $z \in \mathbb{R}^{\mathcal{M}}$, $z \in \operatorname{span}\{y^\phi\}_{\phi\in\mathcal{B}}$. By Claim L1.4, given $z \in \mathbb{R}^{\mathcal{M}}$, we can write $f_z = \sum_{\phi\in\mathcal{B}} \lambda^z_\phi f_\phi$ for some $\{\lambda^z_\phi\}_{\phi\in\Phi} \subseteq \mathbb{R}$. Then for each $\mu \in \mathcal{M}$, $E_{x\sim\mu}[f_z(x)] = z_\mu(1_{i_\mu} \oplus 0_{-i_\mu}) = \sum_{\phi\in\mathcal{B}} \lambda^z_\phi E_{x\sim\mu}[f_\phi(x)]$, and hence $z_\mu = \sum_{\phi\in\mathcal{B}} \lambda^z_\phi y^\phi_\mu$. It follows that $z = \sum_{\phi\in\mathcal{B}} \lambda^z_\phi y^\phi$, and hence $z \in \operatorname{span}\{y^\phi\}_{\phi\in\mathcal{B}}$.

We now complete the proof. Since $\mathcal{B}$ is a basis for $\Gamma(F)$, which is a subspace of $\mathbb{R}^{\dim(\Phi)}$, it has no more than $\dim(\Phi)$ elements; it follows from Claim L1.5 that $\dim(\mathbb{R}^M) = |\mathcal{B}| \leq \dim(\Phi) < |\mathcal{M}|$, a contradiction. ■

Corollary 2. *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 $y \in Y$ be an output. The set of priors*

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

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 $\Gamma^{-1}(\phi^*) \supseteq \Gamma|_{F'}^{-1}(\phi^*)$. Hence

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

The claim then follows immediately from Lemma 1. ■

Proof of Proposition 1 Follows immediately from Corollary 2 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 $y \in Y$ let $X_{\phi,y} \equiv \{x \in X \mid f(x) \neq y \forall f \in \Gamma^{-1}(\phi)\}$. By Proposition 1, each $X_{\phi,y}$ is finite. Since X is convex, it has no isolated points, so it follows that each y, $X \setminus X_{\phi,y}$ is dense in X. Then since u is continuous, for each $y \in Y$ and $a \in A$, $u(X,y,a) \subseteq \mathrm{cl}(u(X \setminus X_{\phi,y}, y, a))$. Hence $\inf_{x\in X} u(x,y,a) \leq \inf_{x \in X\setminus X_{\phi,y}} u(x,y,a)$, and since $X \setminus X_{\phi,y} \subseteq X$, we have

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

Now by definition of $X_{\phi,y}$, for each $y \in Y$ and $a \in A$,

$$\begin{aligned}\{u(x,y,a) \mid x \in X \setminus X_{\phi,y}\} &= \{u(x,f(x),a) \mid f \in \Gamma^{-1}(\phi), x \in X \setminus X_{\phi,y}, f(x) = y\} \\ &\subseteq \{u(x,f(x),a) \mid f \in \Gamma^{-1}(\phi), x \in X\}.\end{aligned}$$

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

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

Taking infima over y yields $\inf_{x\in X, y\in Y} u(x,y,a) \geq \inf_{x\in X, f\in\Gamma^{-1}(\phi)} u(x,f(x),a)$. Then 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. Then taking maxima over A yields $R(\phi|\Gamma) = \underline{R} = \inf_{x\in X, y\in Y} u(x,y,a)$, as desired. ■

Proof of Theorem 4 (Futility of Explanation with Ambiguity Aversion) Fix $\phi \in \Gamma(F)$, and for each $y \in Y$ let $M_{y,\phi}$ be as in (3). By Corollary 2, for each $y \in Y$, $\dim(M_{y,\phi}) \leq \dim(\Phi) < \dim(\mathcal{M})$.

Claim T4.1. $\mathcal{M} \setminus M_{y,\phi}$ is dense in $\mathcal{M}$ (in the weak*-topology): Since $\dim(\mathrm{aff}(M_{y,\phi})) = \dim(M_{y,\phi}) < \dim(\mathcal{M})$, $\mathcal{M} \setminus \mathrm{aff}(M_{y,\phi})$ is nonempty. Given $\mu \in M_{y,\phi}$, choose $\mu' \in \mathcal{M} \setminus \mathrm{aff}(M_{y,\phi})$. Then for each n, $\mu_n = \frac{1}{n}\mu' + (1 - \frac{1}{n})\mu \in \mathcal{M}$ (since $\mathcal{M}$ is convex) but $\mu_n \notin \mathrm{aff}(M_{y,\phi})$ (since if it was, then because $\mu \in \mathrm{aff}(M_{y,\phi})$, we would have to have $\mu' = n\mu_n - (n-1)\mu \in \mathrm{aff}(M_{y,\phi})$)). Since $\mu_n \to_{w^*} \mu$, μ is a limit point of $\mathcal{M} \setminus M_{y,\phi}$; the claim follows.

Since u is continuous, for each $y \in Y$ and $a \in A$, it follows from Claim T4.1 that $\{E_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M}\} \subseteq \mathrm{cl}(\{E_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M} \setminus M_{\phi,y}\})$. Hence $\inf_{\mu\in\mathcal{M}} E_{x\sim\mu}[u(x,y,a)] \leq \inf_{\mu \in \mathcal{M}\setminus M_{\phi,y}} E_{x\sim\mu}[u(x,y,a)]$, and since $\mu \in \mathcal{M} \setminus M_{\phi,y} \subseteq \mathcal{M}$, we have

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

Since u is quadratic, we have $u(x,y,a) = w_0(a) + w_1(a)^\intercal x + w_2(a)^\intercal y$. Then by definition of $M_{\phi,y}$, for each $y \in Y$ and $a \in A$,

$$\begin{aligned}
\{E_{x\sim\mu}[u(x,y,a)] \mid \mu \in \mathcal{M} \setminus M_{\phi,y}\} &= \{w_0(a) + w_1(a)^\intercal E_{x\sim\mu}[x] + w_2(a)^\intercal y \mid f \in \Gamma^{-1}(\phi), \mu \in \mathcal{M} \setminus M_{\phi,y}\} \\
&= \{w_0(a) + w_1(a)^\intercal E_{x\sim\mu}[x] + w_2(a)^\intercal E_{x\sim\mu}[f(x)] \mid f \in \Gamma^{-1}(\phi), \mu \in \mathcal{M} \setminus M_{\phi,y}, E_{x\sim\mu}[f(x)] = y\} \\
&\subseteq \{E_{x\sim\mu}[u(x,f(x),a)] \mid f \in \Gamma^{-1}(\phi), \mu \in \mathcal{M}\}.
\end{aligned}$$

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

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

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

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

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. ■