EXPLANATIONS FOR ROBUST DECISIONS

By

Kai Hao Yang, Nathan Yoder and Alexander K. Zentefis

July 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2552

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Explanations for Robust Decisions*

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

July 21, 2026

Abstract

We consider the problem of explaining a data generating process (DGP) to a decision maker (DM) who cannot understand it. Explanations are information, not approximations: they are useful because they rule out possible DGPs. If the DM maximizes her average payoff, explanations using OLS are robustly useful, and augmenting them with summary statistics makes them more so. Sampling error can destroy these guarantees, but theoretical assumptions linking that error to the DGP can restore them. If the DM is sufficiently ambiguity averse, explanations that are affine in outcomes are not robustly useful, but certificates reporting lower bounds on outcomes are.

JEL classification: C50, D81

Keywords: data-generating processes, decision-making, explanations, least squares

*We thank Arjada Bardhi, Joel Flynn, Mira Frick, Paul Goldsmith-Pinkham, Kevin He, Ryota Iijima, Annie Liang, Omer Tamuz, Akhil Vohra, and Mark Whitmeyer for their very helpful comments, as well as participants at the 2024 Kansas Workshop in Economic Theory, EC'25, 2026 UMiami Theory Day, and 2026 NASMES, and seminar participants at Boston College, Georgia Tech, NYU, Santa Clara Leavey, Pepperdine, and UC Santa Cruz. Yoder acknowledges summer support from the University of Georgia through a Terry-Sanford Research Award. This paper was written with the assistance of GPT, Claude, Gemini, and GitHub Copilot. An extended abstract of this paper (then titled "Explaining Models") appeared in *EC '25: Proceedings of the 26th ACM Conference on Economics and Computation.*

†Yale School of Management, Email: kaihao.yang@yale.edu

‡University of Georgia, John Munro Godfrey Sr. Department of Economics, Email: nathan.yoder@uga.edu

§Leavey School of Business, Santa Clara University, Email: azentefis@scu.edu

1 Introduction

People must often make decisions in environments that are too complicated for them to understand. Policymakers evaluate social programs whose treatment effects vary with individual characteristics in complex ways. Regulators write rules for the deployment of complex artificial intelligence systems without truly knowing how these systems work. How useful to decision makers can intelligible *explanations* of their environments be instead?

We study this question by considering the problem of a decision maker (henceforth DM) who faces a data generating process (DGP) that is too complicated for her to understand. The DGP is a pair of random vectors: *outcomes* Y that specify the DM's payoffs from each of her actions, and *covariates* X (e.g., demographic variables) that the DM may be able to condition her action on. The DGP may be a process found in nature, or it may describe the behavior of a complex artificial system, such as a large AI model responding to prompts.

Because the DGP cannot be understood directly, it must be *explained* to the DM; i.e., mapped onto a small space of reports that she can understand. We call such a map an *explainer* and its output an *explanation*. An explanation may be a statistic, such as the average outcome of each action, or a model, such as a least-squares regression of outcomes on covariates.

We treat explanations as *information*: observing an explanation allows the DM to rule out every DGP that would not have produced it. An explainer thus partitions the space of possible DGPs, and allows the DM to make different choices for DGPs in different elements of that partition. Accordingly, we do not focus on how well an explanation approximates some feature of the DGP. Instead, we are interested in whether the explainer partitions the space of possible DGPs in a way that is useful to the DM by allowing her to make better decisions.

Since the DM does not have a prior over the space of possible DGPs — a space whose elements are too complicated for her to understand — we evaluate explainers by the payoffs that they allow the DM to guarantee. An explainer is *useful for robust decisions* if it allows the DM to guarantee herself a higher payoff, across the DGPs consistent with each explanation, than she could guarantee without an explanation. The most an explainer can achieve is to be *robustly perfect*: to allow the DM to guarantee the payoff she would obtain if she understood the true DGP itself.

Two features of the DM's problem determine which explanations are useful and how useful they are. The first feature is whether she can *target* her actions using the covariates. If, for instance, a policymaker can condition a social program's eligibility on demographic characteristics, we say that she faces a *targeted* problem; if she cannot (e.g., because of legal

constraints or data limitations), we say that her problem is *untargeted*. The second feature is how the DM evaluates her payoff. She may care about her average payoff; e.g., if she is a policymaker maximizing total surplus over a population. She may instead care about her worst-case payoff, as might a regulator concerned with the most harmful response an AI system could produce. Or she may lie in between, evaluating each policy under the least favorable of a set of priors, as in the model of ambiguity aversion of Gilboa and Schmeidler (1989).

The first half of the paper shows that a DM who cares about her average payoff is well served by explanations that are familiar to empirical work. If her problem is untargeted, the average outcome of each action is all she needs: the mean explanation is robustly perfect (Theorem 1), because her payoff from any policy depends on the DGP only through these averages.

But if she faces a targeted problem, the DM can do better if she has information about how outcomes vary with covariates, which a fitted regression reports. We show that an ordinary least squares regression provides her a payoff guarantee for a targeted *policy* (a map from covariates to action probabilities) precisely when the policy lies in the span of its regressors. For instance, a regression of outcomes on a quadratic in income guarantees payoffs for eligibility rules whose treatment probabilities are quadratic in income, but implies nothing about the payoff of an income cutoff, because among the DGPs consistent with the regression, some DGPs make the cutoff perform arbitrarily badly. These payoff guarantees make OLS useful for robust decisions (Theorem 2).

However, the guarantees offered by OLS are not exact. This is because the fitted model only provides information about the relationship between covariates and outcomes, but the DM's payoff from a targeted policy also depends on the distribution of covariates. Reporting information about that distribution — in the form of the means and covariances of the regressors, which appear in standard regression output — alongside the regression coefficients is *more* useful for robust decisions than OLS, because it allows the DM to exactly pin down her payoff from every policy in the span of the regressors (Theorem 3). As the set of regressors grows rich, this *augmented* OLS explanation allows the DM to guarantee herself a payoff that is arbitrarily close to what she could achieve with full knowledge of the DGP (Theorem 4); the fitted model alone, however, does not.

An important caveat to these results is that they concern explanations computed under the same population distribution that determines the DM's average payoff. They apply to an explainer with access to the entire DGP — as when the DGP is an AI model that can be queried at will — but not necessarily to an *estimator* that only has access to a sample from it. Theorem 5 and Theorem 6 show that this distinction is not benign: however small the

sampling error, there are sampling distributions under which the sample mean and the OLS estimator cannot robustly improve the DM's decisions. The problem is not that the estimates are inaccurate — they can be arbitrarily close to the population values — but that the set of DGPs consistent with a sampled report remains too large. Sample-based explanations can therefore support robust decisions only in combination with assumptions about the DGP and/or the form of the sampling error. We identify one such assumption that fully restores our positive results: when selection into the sample depends on the covariates alone, in a known way, reweighting the sample recovers every population explanation together with its guarantees (Proposition 6).

When the DM is ambiguity averse, the mean and OLS explainers fare much worse. Both share a common structure: they are *affine* in the random vector of outcomes determined by the DGP. Theorem 7 shows that explainers with this structure are not useful for robust decisions if the ambiguity-averse DM's set of priors has higher dimension than the space of explanations. Intuitively, given a policy, an ambiguity-averse DM cares about a set of moments of the DGP with the same dimension as her set of priors. If an explainer is affine in the DGP's outcome process, it identifies a set of moments with the same dimension as its range. If the former is higher than the latter, then any explanation must fail to pin down the DM's payoff under some of her priors. In the special case where the DM cares about the worst outcome produced by the DGP, this dimensionality condition is always satisfied (Corollary 5).

These negative results do not rule out the possibility of robustly useful explanation to a sufficiently ambiguity averse DM — they just suggest that affine explainers are not the right tool for the job. We show that *certificates* — explainers which report each action's worst expected payoff from any of the DM's priors, either overall (in an untargeted problem) or conditional on a partition of the space of covariates (in a targeted problem) — are useful for robust decisions, no matter how ambiguity averse the DM is (Theorem 8, Theorem 9).

Overall, the paper's results imply that the right way to explain a data generating process depends on the way that the DGP is evaluated by the DM being explained to. Means and regressions serve a DM whose preferences over DGPs are captured by expected utility, while certificates serve a DM who is ambiguity-averse or cares about the DGP's worst-case outcome.

Outline The remainder of the paper proceeds as follows. Section 2 describes the paper's setting, and discusses its relationship to the literature on model evaluation and statistical decision theory. Section 3 provides the main results when the DM cares about her average payoff, including the analysis of sampling error. Section 4 provides the main results when the DM is ambiguity averse or cares about her worst-case payoff, and introduces certificates.

Section 5 provides a discussion of the paper's findings. Section 6 describes the relationship between our work and several strands of related literature. Section 7 concludes.

2 Setting

Actions and Payoffs A decision maker (henceforth DM) must take an action a in a finite set $A = \{a_1, \ldots, a_M\}$ without knowing the payoff y_a she will receive from taking it.

Covariates and Outcomes A state of the world is $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^K$ is a compact convex set with $\dim(\mathcal{X}) = K$, and $\mathcal{Y} = \mathbb{R}^M$. $x \in \mathcal{X}$ is an exogenous *covariate vector*, while $y \in \mathcal{Y}$ specifies the payoffs that the DM receives after each of her $M \geq 2$ actions.

Data Generating Processes A *data generating process (DGP)* is a pair of bounded measurable random vectors $f = (X, Y)$ on a probability space $([0,1], \mathcal{B}, \mu_0)$, where $X : [0,1] \to \mathcal{X}$ and $Y : [0,1] \to \mathcal{Y}$. Given μ_0, the DGP determines the distribution of states of the world: when $f = (X, Y)$ is the true DGP, the state of the world is given by $(x, y) = (X(\omega), Y(\omega))$ for $\omega \sim \mu_0$. Without loss, we assume that μ_0 has full support.

The DM does not know the true DGP. She only knows that it is some element of the set of *possible* DGPs $F \subseteq \mathcal{X}^{[0,1]} \times \mathcal{Y}^{[0,1]}$. In general, we assume that F is the set of all bounded Borel measurable DGPs $f = (X, Y)$ such that $X : [0,1] \to \mathcal{X}$ is continuous and onto.[1]

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* $x \in \mathcal{X}$. Each outcome $y \in \mathcal{Y} = \mathbb{R}^M = \mathbb{R}^2$ 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.

Example 2 (Frontier AI Regulation)**.** A regulator needs to set policies for the regulation of frontier AI systems by choosing among finitely many rules $a \in A$ (e.g., export controls, security guardrails, or a total ban). Covariates $\mathcal{X} \subseteq \mathbb{R}^K$ denote prompts that may be sent to the AI (e.g., "fold this protein", "fix this code"). An outcome $y \in \mathcal{Y} = \mathbb{R}^M$ describes the net surplus y_a that the AI's response generates when subject to each rule a.

Example 3 (Training a Model)**.** An AI lab is training a language model that will output a token $a \in A$ at position $t + 1$ after reading a length-t sequence of tokens $x \in \mathcal{X}$. Each outcome y describes the utility that a user will get from each possible next token, so that y_a is the utility from token a.

[1] Formally, $F := \{(X, Y) \in C([0,1], \mathcal{X}) \times B_b([0,1])^M \mid X([0,1]) = \mathcal{X}\}$.

Example 4 (Loan Origination)**.** A bank must decide what interest rate $a \in A$ to charge loan applicants. Covariates $x \in \mathcal{X}$ denote applicant characteristics (e.g., credit score, income, employment history) and outcomes y_a denote the net present value of the stream of payments that the applicant makes under interest rate a.

Example 5 (Asset Allocation)**.** An asset manager must decide the proportion of her portfolio to allocate to each asset $a \in A$. Covariates $x \in \mathcal{X}$ denote public market conditions (e.g., recent history of asset prices) and outcomes y_a denote the return of asset a over the following period.

Explanations, Statistics, and Models Data generating processes are, in general, too complicated for the DM to understand. Instead, the DGP must be *explained* to her using a map Γ from the set F of possible DGPs to a finite-dimensional space $\Phi = \Gamma(F)$. We call the map Γ an *explainer* and Φ its *space of explanations*. Two types of explanations are of particular interest.

- **Models.** The DM may want the relationship between covariates and outcomes captured by the DGP to be explained to her using a map from $\mathcal{X}$ to $\mathcal{Y}$. We say that Γ is a *model* if Φ is an n-dimensional linear subspace of $\mathcal{Y}^{\mathcal{X}}$; e.g., the set of nth degree polynomials in x.

  Example 6 (The OLS Explainer)**.** Let $\mu \in \Delta([0,1])$ be a probability measure with full support on $[0,1]$. Given a finite-dimensional linear subspace of continuous *regressors* $\hat{\Phi} \subseteq C(\mathcal{X})$, any covariate process $X : [0,1] \to \mathcal{X}$ induces a subspace $\Psi_X := \{\phi \circ X : \phi \in \hat{\Phi}^M\} \leq L^2([0,1], \mu; \mathcal{Y})$ of achievable approximations to Y using regression models in $\overline{\Phi} := \hat{\Phi}^M$. The *ordinary least squares (OLS) explainer* under μ, $\overline{\Gamma}_\mu$, returns the regression model that minimizes the mean-squared error of this approximation.

  Formally, $\overline{\Gamma}_\mu(f) \in \overline{\Phi}$ is defined by the orthogonal projection $P^\mu_{\Psi_X}$ onto Ψ_X:

$$\overline{\Gamma}_\mu(f) \circ X = P^\mu_{\Psi_X}(Y).$$

  Equivalently, $\overline{\Gamma}_\mu(f)$ solves the standard least-squares problem:

$$\overline{\Gamma}_\mu(f) = \operatorname*{argmin}_{\phi \in \overline{\Phi}} \int_0^1 \|Y(\omega) - \phi(X(\omega))\|^2 \, \mathrm{d}\mu(\omega),$$

  where $\|\cdot\|$ is the Euclidean norm on $\mathcal{Y}$.[2] When $\mu = \mu_0$, we write $\overline{\Gamma} := \overline{\Gamma}_{\mu_0}$. We assume

[2]The projected random variable $P^\mu_{\Psi_X}(Y)$ is uniquely defined in $L^2([0,1], \mu; \mathcal{Y})$. The corresponding regres-

that the space of regressors contains a constant: $\mathbf{1} \in \hat{\Phi}$, and hence $\overline{\Phi}$ contains the constant functions.

- **Statistics.** The DM may want the DGP to be explained to her using statistics about it (e.g., its moments). We say that Γ is a *statistic* if $\Phi = \mathbb{R}^n$.

  Example 7 (The Mean Explainer)**.** While the OLS explainer describes how outcomes vary with covariates, decision makers who cannot tailor their actions to the covariates are primarily concerned with the *unconditional* average outcome. For any distribution $\mu \in \Delta([0,1])$, the *mean explainer* $\mathbb{e}_\mu$ simply returns the expected outcome vector:

$$\mathbb{e}_\mu(f) := \int_0^1 Y(\omega)\,\mathrm{d}\mu(\omega) \in \mathbb{R}^M.$$

  When $\mu = \mu_0$, we write $\mathbb{e} := \mathbb{e}_{\mu_0}$ to denote the population mean explainer.

Decision Problems and Policies Henceforth, we refer to a *decision problem* by a tuple (Π, V), where Π is the set of *policies* that the DM can choose from, and $V : F \times \Pi \to \mathbb{R}$ describes the DM's indirect utility from choosing policy $\pi \in \Pi$ when the true DGP is $f \in F$.

We consider two types of policies that may be available to the DM.

- In some problems, the DM can observe the covariate vector x and condition her action on it. For example, a policymaker may implement a means-tested treatment, a regulator may require self-driving cars to obey different rules in different conditions, or an AI lab may train its language model to output different tokens after different sequences. In such *targeted* decision problems, Π is the space of *targeted policies*, or Borel-measurable maps $\pi : \mathcal{X} \to \Delta(A)$, where $\Delta(A)$ denotes the set of probability distributions over actions.

- In other problems, the DM either cannot observe the covariate vector x or cannot condition her action on it. E.g., the U.S. Constitution or the Civil Rights Act may prevent a policymaker from targeting different treatments at individuals with different characteristics, or technical limitations may prevent a regulator from tailoring self-driving car regulation to specific conditions. In such *untargeted* decision problems, Π is simply the space $\Delta(A)$ of probability distributions over actions.

sion model $\phi \in \overline{\Phi}$ is also unique provided the regressors are not multicollinear when applied to the covariate process X; i.e., if whenever $\hat{\phi} \in \overline{\Phi}$ satisfies $\hat{\phi}(X(\omega)) = 0$ for μ-almost every $\omega \in [0,1]$, we have $\hat{\phi} = 0$. This follows from the fact that the regressors are continuous and (since X is continuous and onto, and μ has full support on $[0,1]$) the pushforward measure $X_{\#}\mu$ has full support on $\mathcal{X}$.

Decision Rules The explanation $\phi \in \Phi$ that is provided by an explainer Γ may not precisely identify the true DGP. But it does allow the DM to rule out all data generating processes that are not *consistent* with ϕ, i.e., all DGPs that are not in

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

The explainer's function is thus to partition the set F of possible DGPs into the preimages $\{\Gamma^{-1}(\phi)\}_{\phi\in\Phi}$, and allow the decision maker to choose different policies on different elements of the partition: Explanations serve as *information* about the true DGP, not approximations. The decision maker's policy choices are then captured by a *decision rule* $\alpha : \Phi \to \Pi$.

The relationship between a data generating process f, an explainer Γ, a decision rule α, and a policy π is summarized in Figure 1 below.

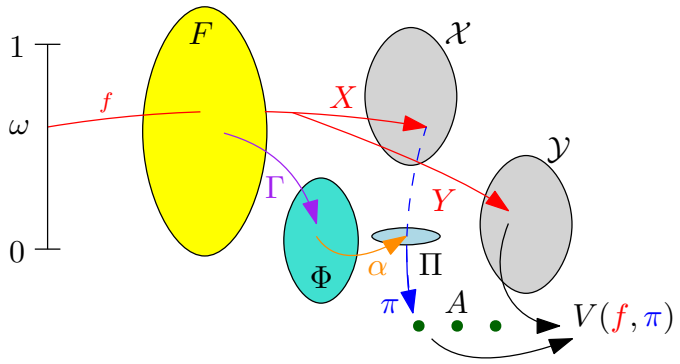

Figure 1: DGPs, explainers, decision rules, and policies. The figure depicts the space F of possible true DGPs $f = (X, Y)$, which are pairs of random vectors taking values in $\mathcal{X}$ and $\mathcal{Y}$; an explainer Γ that maps the space F of possible true DGPs to the space Φ of explanations; and a decision rule α that maps the space Φ of explanations to the set Π of policies, which are either distributions over the set of actions A or maps from the set of covariates $\mathcal{X}$ to such distributions, depending on whether the problem is targeted or untargeted.

Example 1 (Treatment Effects), continued. When treatment decisions can depend on covariates (i.e., when her problem is *targeted*), our policymaker might observe the fitted model $\overline{\Gamma}(f)$ from a regression of the potential outcome vector on a polynomial in the covariates. Then, she will choose a policy π that determines the probability with which an individual with covariates x will be treated. Alternatively, if the treatment decision cannot be conditioned on the covariates (i.e., when her problem is *untargeted*), she might obtain only an average treatment effect $\mathbb{e}(f)$ for the population, and then choose the fraction of the population that will receive the treatment. In either case, the policymaker's choice of policy will depend on the explanation Γ via a decision rule α.

Example 2 (AI Regulation), continued. When regulations can depend on the prompt (e.g., guardrails triggered only by sensitive cybersecurity questions), a regulator might ask for a fitted model $\Gamma(f)$ describing the relationship between the prompt x and the usefulness or harm to society caused by its output when the model is subject to different kinds of safeguards. Then, he will choose a policy π describing which safeguards to apply to which kinds of prompts. If this is not possible — e.g., if interventions cannot be applied at the prompt level, and the only tools available are export controls — the regulator might simply rely on a statistic describing the model's responses (e.g., adversarial robustness evaluations as in Anthropic PBC (2026), p. 68). The map from these explanations to policies then describes the regulator's decision rule α.

Example 3 (Training a Model), continued. When an AI lab is training a language model through reinforcement learning, they may use a statistical model to estimate the relationship between the input string x and the utility y_a of each possible output token a (a *reward model*). Then, they can use this model to find a targeted policy $\pi : \mathcal{X} \to \Delta(A)$ that gives a high value of the reward model. We denote the map from reward models to chosen policies as their decision rule α.

2.1 Robust Explanation

Our analysis focuses on evaluating various explainers according to the usefulness of the information they provide about the DGP. Since the DM cannot discriminate between the DGPs that are consistent with a given explanation, we evaluate an explainer by whether it allows the DM to improve his payoff robustly across all of them.

Robust Usefulness We say that explanation with Γ is *useful for robust decisions* if it allows the DM to do better than committing to a single policy — that is, if there is a decision rule $\alpha : \Phi \to \Pi$ such that for every policy $\pi \in \Pi$:

$$\inf_{f \in \Gamma^{-1}(\phi)} V(f, \alpha(\phi)) \geq \inf_{f \in \Gamma^{-1}(\phi)} V(f, \pi) \quad \text{for all } \phi \in \Phi, \text{ and} \tag{1}$$

$$\inf_{f \in \Gamma^{-1}(\phi)} V(f, \alpha(\phi)) > \inf_{f \in \Gamma^{-1}(\phi)} V(f, \pi) \quad \text{for some } \phi \in \Phi. \tag{2}$$

That is, no matter what fixed policy π the DM would otherwise have chosen, explaining the DGP with Γ lets him attain a payoff guarantee — across the DGPs consistent with each explanation — that is always at least as good and sometimes strictly better. Otherwise, we say that Γ is *useless for robust decisions*.

Conversely, explanation cannot make the DM better off than understanding the true DGP. We say that explanation with Γ is *robustly perfect* if it allows the DM to guarantee himself this full-information benchmark: if there exists a decision rule α such that

$$\overline{V}(f) := \sup_{\pi \in \Pi} V(f, \pi) = V(f, \alpha(\Gamma(f))) \text{ for all } f \in F.$$

Identification Explanation works by partitioning the space of DGPs. To be useful, this partition must group DGPs with similar payoffs from at least one policy π into the same element $\Gamma^{-1}(\phi)$. When it does — and so

$$\inf_{f \in \Gamma^{-1}(\phi)} V(f, \pi) > \underline{V}(\pi) := \inf_{f \in F} V(f, \pi),$$

we say that ϕ *identifies* π. When every $\phi \in \Phi$ identifies π, we just say that π is identified by Γ.

This identification is most useful when the payoff from π is not only similar across DGPs in $\Gamma^{-1}(\phi)$, but *the same*. We say that ϕ *exactly identifies* π if $V(f, \pi) = V(g, \pi) > \underline{V}(\pi)$ for all $f, g \in \Gamma^{-1}(\phi)$. When every $\phi \in \Phi$ exactly identifies π, we just say that π is exactly identified by Γ. Obviously,

Lemma 1. *If Γ exactly identifies every $\pi \in \Pi$, then Γ is robustly perfect.*

2.2 Discussion

Our setting and central question are both tightly linked to the literature on the evaluation of theoretical models (e.g., Fudenberg and Liang 2020; Montiel Olea, Ortoleva, Pai and Prat 2022; Fudenberg, Kleinberg, Liang and Mullainathan 2022; Andrews, Fudenberg, Liang and Wu 2023; Fudenberg, Gao and Liang 2026). Like those papers, we consider methods for simplifying a complex true relationship between inputs $\mathcal{X}$ and outputs $\mathcal{Y}$ using a map from the former to the latter — what they call a *prediction rule* and we call an *explanation* from a model. 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 (the space of regressors in our framework). Our key departure is the following: In their settings, prediction rules are *approximations*, and are evaluated by how close they are to some feature of the data generating process; in our setting, explanations are *information*, and are evaluated by how useful they are to a decision maker.

Our setup is also related to that considered in statistical decision theory (e.g., Wald 1949; Savage 1951). There, the statistician observes *data* from a *sampling distribution* that depends

on the *state of the world*. Our baseline analysis abstracts from sampling error, and focuses on explanation instead: in our model, the DM observes an *explanation* from a deterministic *explainer* that takes the *true DGP* as its argument. As suggested by Wald (1949), we consider both *average* (Section 3) and *worst case* (Section 4) payoffs. However, a key difference in our setting is that the average or worst case is not over DGPs (which play the same role as "states" in statistical decision theory, and over which we always consider the worst case) but over the underlying probability space itself.

Within this literature, the most closely related paper is Cheng (2026), who studies a decision maker that observes a finite sequence of realizations from an unknown data generating process and chooses an *act* that maps future realizations to payoffs. As in our setting, his DM focuses on the worst-case expected payoff over a subset of DGPs that is determined by his information. This resembles the untargeted utilitarian problem with sampling error in Section 3.3, but his question is subtly different: In his setting, the acts are fixed, and the unknown DGP is the distribution on the underlying probability space, while in our setting, the DGP governs the acts themselves (i.e., the map from the underlying probability space to payoffs). Like us, Cheng (2026) derives an impossibility result showing that when the mappings from the underlying probability space to payoffs are unrestricted (in his language, acts; in our language, DGPs), no finite-dimensional report can guarantee improved decisions under sampling error. However, he pursues a different remedy: instead of a restriction based on the relationship between sampling error and an observable covariate, he considers one based on monotonicity.

3 Utilitarian Decision Problems

We first focus on decision problems in which the decision maker's preferences over policies and data generating processes have an expected utility representation. Specifically, we suppose that the DM's indirect utility from a DGP $f = (X, Y)$ is given by the *expectation* of his payoff with respect to the laws of X and Y. 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 labs may only care about the average performance of AI models they regulate or train.

In particular, suppose the true DGP is $f = (X, Y)$. Then in an untargeted utilitarian decision problem, a policy $\pi \in \Delta(A)$ yields indirect utility

$$U(f, \pi) := \mathbb{E}_{\omega \sim \mu_0}[\pi \cdot Y(\omega)]$$

Similarly, in a targeted utilitarian decision problem, a policy $\pi : \mathcal{X} \to \Delta(A)$ yields indirect utility

$$U(f, \pi) := \mathbb{E}_{\omega \sim \mu_0} [\pi(X(\omega)) \cdot Y(\omega)]$$

3.1 Untargeted Decision Problems

In untargeted decision problems, the mean explainer $\mathbb{e}$ gives a utilitarian DM a *robust explanation* that is *always perfect*: it *always* (for any possible true model) gives an explanation that allows the DM to make a decision that is *robustly* (across all DGPs that are consistent with the explanation) better, and in fact *perfect* (in the sense that the DM could do no better by knowing the true model).

Theorem 1 (Mean Explanations Are Robustly Perfect)**.** *If the DM is utilitarian and faces an untargeted decision problem, then explanation with the mean explainer $\mathbb{e}$ is robustly perfect. In particular, the naïve decision rule $\alpha^*(\phi) \in \arg\max_{\pi \in \Delta(A)} \pi \cdot \phi$ always achieves the full-information benchmark: For all $f \in F$,*

$$U(f, \alpha^*(\mathbb{e}(f))) = \overline{U}(f).$$

Theorem 1 is deliberately immediate. Recall that in an untargeted problem, a utilitarian DM only cares about the DGP's expected outcome $\mathbb{E}[Y]$. This is precisely the information about the DGP preserved by the mean explainer: for all DGPs f, $\hat{f} \in \mathbb{e}^{-1}(\phi)$ and policies $\pi \in \Delta(A)$, we have $U(f, \pi) = U(\hat{f}, \pi) = \pi \cdot \phi$. Thus, a utilitarian DM need not be sophisticated to get the full value of robust explanation from $\mathbb{e}$. Instead, they can achieve their first-best value $\overline{U}(f)$ by naïvely choosing a policy (or, since U is linear in π, an action) as if the explanation $\phi = \mathbb{e}(f)$ was the true model.

3.2 Targeted Decision Problems

Explanation is more challenging in targeted problems. Once the DM can condition her action on the covariate vector x, the mean explainer is no longer sufficient to achieve the full-information benchmark, because (as we show in Theorem 3) it only pins down the payoffs from constant policies. Instead, the explainer must describe how the outcome y varies with x in order to be useful for robust choice among targeted policies.

A natural benchmark that provides this information is the OLS explainer $\overline{\Gamma}$. We thus begin by asking which targeted policies give the DM a payoff that is pinned down by an OLS explanation. Given the space of regression models $\overline{\Phi} = \hat{\Phi}^M$, define the set of *OLS-identified policies* as

$$\Pi_{\overline{\Phi}} := \overline{\Phi} \cap \Delta(A)^{\mathcal{X}}.$$

These are the policies that lie in the space of regression models $\overline{\Phi}$. The robust content of an explanation from OLS is exactly the part of the targeted policy problem that remains inside this identified class.

Proposition 1. *Let $\phi \in \overline{\Phi}$ and $\pi \in \Pi_{\overline{\Phi}}$. Then in a targeted utilitarian decision problem, ϕ identifies π, with*

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi)} U(f, \pi) = \inf_{x \in \mathcal{X}} \pi(x) \cdot \phi(x).$$

Intuitively, OLS identifies the payoff of policies in the span of the regressors given a fixed distribution of covariates, but not the distribution of covariates itself. The robust guarantee it provides is thus given by the worst-case payoff over covariate distributions, which is approached as the distribution converges to a point mass.

Policies outside $\Pi_{\overline{\Phi}}$ fare much worse. Once a policy leaves the space of regressors, its expected payoff can be decreased arbitrarily by perturbing the DGP along directions that do not affect the OLS explanation.

Proposition 2. *Let $\phi \in \overline{\Phi}$ and $\pi \notin \Pi_{\overline{\Phi}}$. Then in a targeted utilitarian decision problem, ϕ does not identify π.*

The reason that policies in $\overline{\Phi}$ are identified by the OLS explainer (Proposition 1), and those outside this class are not (Proposition 2), is orthogonality. Write the outcome process as $Y = \phi(X) + R$, where ϕ is the OLS explanation of $f = (X, Y)$ and the residual process R is orthogonal to every OLS explanation: $R \in \Phi^{\perp}$. This is the only restriction on R implied by the model's consistency with ϕ. Hence, if the DM chooses a policy π that is not orthogonal to $\Phi^{\perp}$ — i.e., one not in Φ — there is some DGP consistent with ϕ for which π produces an arbitrarily low payoff.

Propositions 1 and 2 characterize an optimal decision rule for the DM. If she wants to make decisions that are robust to model uncertainty, a utilitarian decision maker who observes an explanation from OLS should maximize her worst-case payoff over identified policies:

$$\alpha^{\mathrm{OLS}}(\phi) \in \operatorname*{argmax}_{\pi \in \Pi_{\overline{\Phi}}} \inf_{x \in \mathcal{X}} \pi(x) \cdot \phi(x).$$

This provides the best robust guarantee that OLS can support.

Theorem 2 (OLS Is Robustly Useful)**.** *Explanation with the OLS explainer $\overline{\Gamma}$ is useful for robust decisions in targeted utilitarian problems. In particular, for every targeted policy $\pi : \mathcal{X} \to \Delta(A)$ and every explanation $\phi \in \overline{\Phi}$,*

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi)} U(f, \alpha^{\mathrm{OLS}}(\phi)) \geq \inf_{f \in \overline{\Gamma}^{-1}(\phi)} U(f, \pi),$$

with strict inequality for some ϕ.

Example 1 revisited. A policymaker is deciding the eligibility criteria for a new government program. She observes an OLS regression of the vector $y = (y_0, y_1)$ of the program's potential outcomes on a space $\hat{\Phi}$ of regressors (functions of demographic characteristics x) with basis b. After observing the fitted regression model $\phi \in \overline{\Phi}$, she can guarantee that any policy $\pi \in \Pi_{\overline{\Phi}}$ will create total surplus of at least $\inf_{x \in \mathcal{X}} \pi(x) \cdot \phi(x)$.

Theorem 2 shows that OLS is robustly useful for targeted utilitarian problems. But unlike the mean explainer in untargeted problems, it is not robustly perfect: the lower-envelope characterization in Proposition 1 does not generally coincide with the full-information benchmark.

A key reason for this gap is that while each policy in $\Pi_{\overline{\Phi}}$ is identified by the OLS explainer, none of them is identified exactly. While OLS reports information about the relationship between x and y (in the form of the fitted model ϕ), it does not provide the DM with information about the distribution of x itself. As it turns out, the missing object needed for exact identification is the *Gram matrix* of a basis for the space of regressors.

Given a basis $b = (b_1, \ldots, b_{\dim(\hat{\Phi})})$ of $\hat{\Phi}$, define the *Gram matrix of b under f* as

$$\Sigma^b(f) := \mathbb{E}_{\omega \sim \mu_0} \left[b(X(\omega)) b(X(\omega))^\top \right].$$

The *augmented OLS explainer* $\widetilde{\Gamma}^b : F \to \widetilde{\Phi}$ for b reports this object along with the OLS explanation:

$$\widetilde{\Phi} := \overline{\Phi} \times \mathbb{R}^{\dim(\hat{\Phi}) \times \dim(\hat{\Phi})}, \qquad \widetilde{\Gamma}^b(f) := (\overline{\Gamma}(f), \Sigma^b(f)),$$

Augmented OLS only requires information that is typically reported in a standard regression output: $\Sigma^b(f)$ can be constructed from the means and covariances of the regressors.[3]

Observe that any OLS explanation $\overline{\Gamma}(f)$ can be written as $\overline{\Gamma}(f) = Bb(x)$, and any OLS-identified policy $\pi \in \Pi_{\overline{\Phi}}$ can be written as $\pi(x) = Gb(x)$, for some $M \times \dim(\hat{\Phi})$ matrices B and G. Proposition 3 shows that in a targeted utilitarian decision problem, augmenting the OLS explanation with $\Sigma^b(f)$ makes its identification of π exact, and pins down the DM's payoff as the *trace* of the product of these three matrices.

Proposition 3. *Let $(Bb, \Sigma) \in \widetilde{\Phi}$, and let $\pi = Gb \in \Pi_{\overline{\Phi}}$, for some $M \times \dim(\hat{\Phi})$ matrices B and G. Then in a targeted utilitarian decision problem, (Bb, Σ) exactly identifies π, with*

$$U(f, \pi) = \operatorname{tr}(G \Sigma B^\top) \text{ for all } f \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma). \tag{3}$$

[3] Observe that $\Sigma^b(f) = \operatorname{Cov}_{\omega \sim \mu_0}(b(X(\omega))) + \mathbb{E}_{\omega \sim \mu_0}[b(X(\omega))] \mathbb{E}_{\omega \sim \mu_0}[b(X(\omega))]^\top$.

The augmented OLS explainer improves on the robust payoff guarantee from Proposition 1 because it provides information (in the form of Σ^b) about the distribution of x. Theorem 3 shows that this information can change which identified policy has the *best* guarantee, making augmented OLS more useful for robust decision-making than the mean or OLS explainers. Relative to the former, augmented OLS identifies a larger set of policies; relative to the latter, it identifies the same set of policies, but does so exactly.

Theorem 3 (Augmented OLS Identifies More, Is More Informative, and Is More Useful)**.** *Suppose $\hat{\Phi}$ contains a nonconstant function. Then in a targeted utilitarian decision problem:*

(i) (Identification.) *$\widetilde{\Gamma}^b$ identifies strictly more policies than $\mathbb{e}$, and identifies the same policies as $\overline{\Gamma}$, but more precisely:*

- *π is identified by $\widetilde{\Gamma}^b$ if and only if $\pi \in \Pi_{\overline{\Phi}}$; this identification is exact.*
- *π is identified by $\overline{\Gamma}$ if and only if $\pi \in \Pi_{\overline{\Phi}}$; this identification is not exact.*
- *π is identified by $\mathbb{e}$ if and only if π is constant (and thus $\pi \in \Pi_{\overline{\Phi}}$); this identification is exact.*

(ii) (Informativeness.) *Both the OLS and mean explainers are garblings of augmented OLS: there exist maps h_1, h_2 such that $\overline{\Gamma} = h_1 \circ \widetilde{\Gamma}^b$ and $\mathbb{e} = h_2 \circ \widetilde{\Gamma}^b$.*

(iii) (Robust usefulness.) *$\widetilde{\Gamma}^b$ is more useful for robust decisions than $\overline{\Gamma}$ or $\mathbb{e}$: There is a decision rule $\alpha^* : \widetilde{\Phi} \to \Pi$ such that, for each $\widetilde{\Gamma} \in \{\overline{\Gamma}, \mathbb{e}\}$ and every decision rule $\tilde{\alpha} : \widetilde{\Gamma}(F) \to \Pi$,*

$$\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\phi)} U(f, \alpha^*(\phi)) \geq \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\phi)} U(f, \tilde{\alpha}(\widetilde{\Gamma}(f))) \quad \text{for all } \phi \in \widetilde{\Phi}, \text{ and} \tag{4}$$

$$\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\phi)} U(f, \alpha^*(\phi)) > \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\phi)} U(f, \tilde{\alpha}(\widetilde{\Gamma}(f))) \quad \text{for some } \phi \in \widetilde{\Phi}. \tag{5}$$

Parts (ii) and (iii) rank the augmented OLS explainer above both alternatives, in two complementary senses. Part (ii) is an *informativeness* ranking: because the OLS fit and the mean are both functions of the augmented explanation, $\widetilde{\Gamma}^b$ is more informative than either $\overline{\Gamma}$ or $\mathbb{e}$ in the sense of Blackwell (1953) (each is a garbling of $\widetilde{\Gamma}^b$). This ranking guarantees that the better explainer cannot be *less* robustly useful (in the sense of (1) and (2)) in any decision problem. But on its own, it does not ensure that the higher-ranked explainer is actually *more* robustly useful: After all, any explainer is a garbling of the identity map, and an uninformative explainer is a garbling of any explainer, but as we showed in Theorem 1 and show later in Theorems 5, 6, and 7, robustly perfect and robustly useless explainers exist.

Part (iii) proves that this sharper ranking holds in the targeted utilitarian problem: whatever decision rule the DM would have used under an OLS or mean explanation, augmented OLS lets him guarantee a payoff that is always at least as high, and strictly higher for some explanations. Thus, $\widetilde{\Gamma}^b$ is not merely more informative than $\overline{\Gamma}$ and $\mathbb{e}$, but *more useful for robust decisions.*

Asymptotic Robust Perfection

Augmentation closes part of the gap between the robust guarantee offered by OLS and the full-information benchmark: Augmented OLS exactly identifies the payoff from every policy in $\Pi_{\overline{\Phi}}$, rather than just bounding it below by its lower envelope. But augmented OLS is still not robustly perfect. The best guarantee it can support, $\sup_{\pi \in \Pi_{\overline{\Phi}}} U(f, \pi)$, can still fall short of the full-information benchmark $\overline{U}(f)$, because the identified class $\Pi_{\overline{\Phi}}$ need not contain the policy that is optimal with full information. The remaining gap is thus one of approximation: as the space of regressors $\hat{\Phi}$ becomes large, can the full-information benchmark eventually be approximated using policies from $\Pi_{\overline{\Phi}}$?

When the regressors are polynomials, the answer is yes: Theorem 4 shows that as the degree of those polynomials increases, the identified policies become rich enough to recover the first-best asymptotically. Formally, we say that in a decision problem (V, Π), explanation with a sequence of explainers $\{\Gamma_n\}_{n=1}^{\infty}$ is *asymptotically robustly perfect* if there exist decision rules α_n such that for every $f \in F$,

$$\inf_{\hat{f} \in \Gamma_n^{-1}(\Gamma_n(f))} V(\hat{f}, \alpha_n(\Gamma_n(f))) \to \overline{V}(f).$$

Theorem 4 (Augmented OLS Is Asymptotically Robustly Perfect)**.** *For each $n \geq 1$, let $\widetilde{\Gamma}_n^b$ be an augmented OLS explainer with regressor space $\Phi_n := \hat{\Phi}_n^M$, where $\hat{\Phi}_n$ is the space of polynomials on $\mathcal{X}$ of total degree at most n. Then in a targeted utilitarian problem, the sequence $\{\widetilde{\Gamma}_n^b\}_{n=1}^{\infty}$ is asymptotically robustly perfect. In particular, if*

$$\alpha_n^b(Bb, \Sigma) \in \operatorname*{argmax}_{G : Gb \in \Pi_{\overline{\Phi}_n}} \operatorname{tr}(G \Sigma B^\top),$$

then for every $f \in F$,

$$\inf_{\hat{f} \in (\widetilde{\Gamma}_n^b)^{-1}(\widetilde{\Gamma}_n^b(f))} U(\hat{f}, \alpha_n^b(\widetilde{\Gamma}_n^b(f))) \to \overline{U}(f).$$

One might ask whether augmentation with the Gram matrix $\Sigma^b(f)$ is necessary to achieve the full-information benchmark in the limit, or whether enlarging the space of regressors is enough on its own. The answer is no: Proposition 4 shows by counterexample that the gap

between the robust payoff guarantee offered by OLS and the full-information benchmark can be persistently large, regardless of the space of regressors. Hence, OLS alone is not asymptotically robustly perfect: augmentation is necessary to achieve the full-information benchmark in the limit.

Proposition 4. *Suppose $\mathcal{X} = [0,1]$, μ_0 is uniform, and $A = \{1,2\}$. Then in a targeted utilitarian decision problem, explanation with $\overline{\Gamma}$ is not robustly perfect. In particular, there exists a DGP $f^\dagger \in F$ such that regardless of the space of regressors $\hat{\Phi} \subseteq C([0,1])$ used in the OLS explainer $\overline{\Gamma}$,*

$$\sup_{\pi \in \Pi} \inf_{\hat{f} \in \overline{\Gamma}^{-1}(\overline{\Gamma}(f^\dagger))} U(\hat{f}, \pi) \leq \frac{1}{2} < \overline{U}(f^\dagger) = \frac{3}{4}.$$

Corollary 1. *Suppose $\mathcal{X} = [0,1]$, μ_0 is uniform, $A = \{1,2\}$, and for each n, let $\overline{\Gamma}_n$ be the OLS explainer with the space of explanations $\overline{\Phi}_n = \hat{\Phi}_n^2$, where $\hat{\Phi}_n$ is an increasing sequence of finite-dimensional subspaces of $C([0,1])$ that each contain the constant functions. Then in a targeted utilitarian decision problem, explanation with $\{\overline{\Gamma}_n\}_{n=1}^\infty$ is not asymptotically robustly perfect.*

The asymptotic perfection of augmented OLS and the persistent imperfection of unaugmented OLS are two instances of a more general approximation principle. As we note in Lemma 1, if an explainer exactly identifies every policy in Π, then it is robustly perfect. Proposition 5 provides an asymptotic analogue: if each explainer in a sequence $\{\Gamma_n\}_{n=1}^\infty$ exactly identifies every policy in some class Π_n^*, and the classes $\{\Pi_n^*\}_{n=1}^\infty$ asymptotically approximate the first-best, then $\{\Gamma_n\}_{n=1}^\infty$ is asymptotically robustly perfect.

Proposition 5 (Payoff Sufficiency)**.** *Let $\{\Gamma_n\}_{n\geq 1}$ be a sequence of explainers, and for each n, let $\Pi_n^* \subseteq \Delta(A)^{\mathcal{X}}$ be a class of policies that is exactly identified by Γ_n. If*

$$\sup_{\pi \in \Pi_n^*} U(f, \pi) \to \overline{U}(f) \quad \text{for every } f \in F,$$

then $\{\Gamma_n\}_{n=1}^\infty$ is asymptotically robustly perfect.

Proposition 5 decomposes asymptotic robust perfection into two components: (a) exact identification of a class of policies Π_n^*, and (b) asymptotic richness of Π_n^* that allows approximation of the full-information benchmark. With an expanding set of polynomial regressors, augmented OLS satisfies (a) by Theorem 3 with $\Pi_n^* = \Pi_{\overline{\Phi}_n}$ and (b) by Bernstein approximation. Plain OLS, in contrast, fails (a): as Theorem 3 shows, it does not identify any policy exactly, so no enrichment of the regressor space can close the gap in Proposition 4. However, these conditions are satisfied by other explainers, such as cross-moment explainers reporting $\mathbb{E}[Y b_n(X)^\top]$ directly. Thus, augmented OLS is not the unique asymptotically robustly

perfect explainer. It is, however, the canonical way of using OLS to achieve asymptotic robust perfection: it identifies the same class of policies $\Pi_{\overline{\Phi}}$ as OLS (but does so exactly), and requires only regression coefficients and summary statistics about the regressors.

Remark (Comparing Convergence Rates). The payoff sufficiency principle has a useful implication for comparing convergence rates among asymptotically robustly perfect explainers. With a fixed basis b_n, the augmented OLS explainer $\widetilde{\Gamma}^b_n$ and the cross-moment explainer $\Gamma(X,Y) = \mathbb{E}_{\omega\sim\mu_0}[Y(\omega)b_n(X(\omega))^\top]$ exactly identify the same set of policies $\Pi_{\overline{\Phi}_n}$, and thus allow the DM to obtain the same value $\sup_{\pi\in\Pi_{\overline{\Phi}_n}} U(f,\pi)$ from any given DGP $f \in F$. Hence, their payoff guarantees converge to the full-information benchmark at identical rates. More generally, the same is true of any sequences of explainers that exactly identify the same policies: the rate of convergence is determined entirely by the approximation properties of Π^*_n for the first-best policy for f.

Example 5 revisited. An asset manager observes an OLS regression of the vector y of asset returns on a polynomial of degree n in market conditions x. With just the fitted regression model $\overline{\Gamma}(f) \in \overline{\Phi}$, she can guarantee that any asset allocation strategy (a targeted policy) $\pi : \mathcal{X} \to \Delta(A)$ in the space of regression models $\overline{\Phi}$ will yield an expected payoff of at least $\inf_{x\in\mathcal{X}} \pi(x)\cdot\phi(x)$. When she also has access to the Gram matrix $\Sigma^b(f)$ of a vector b of degree-n polynomials in the regressors (or equivalently, their covariance matrix and means), the payoffs from these allocation strategies are exactly pinned down by $\mathrm{tr}(G\Sigma^b(f)B^\top)$, where $\pi = Gb$ and B is the matrix of coefficients from the fitted model $\overline{\Gamma}(f)$. As n becomes large, the space of such identified policies becomes rich enough for the asset manager to achieve a guaranteed return that is almost as good as she could get if she precisely understood the true relationship between market conditions and returns.

3.3 Sampling Error

Theorem 1 suggests that explaining the DGP using a statistic can be very useful for a utilitarian decision maker who is concerned with their *average* payoff over all inputs x. This, however, relies on the fact that the distribution used to construct the mean is precisely the distribution μ_0 that is relevant for the DM's expected payoff. But in the real world, decision makers often only have access to statistics about a *sample* rather than the entire *population*: that is, rather than computing the mean of Y under the *population* distribution μ_0, practitioners compute it under a *sampling* distribution $\mu \in \Delta([0,1])$. This may be due to lack of data (e.g., when f describes the joint distribution of demographic variables and treatment effects, as in Example 1) or to economize on computational resources (e.g., when f is an AI model, as in Example 2).

Theorem 5 shows that the robust explanation offered by the mean explainer is not robust to even small differences in these distributions — i.e., to even small *sampling error*.

Theorem 5 (Sample Means Are Not Robustly Useful)**.** *In an untargeted utilitarian decision problem, the robust usefulness of the mean explainer is not robust to sampling error: for any weak-* open $\mathcal{M} \subseteq \Delta([0,1])$ with $\mu_0 \in \mathcal{M}$, there exists $\mu \in \mathcal{M}$ such that explanation with $\mathbb{e}_\mu$ is useless for robust decisions. That is, for every $\phi \in \Phi$ and every $\pi \in \Delta(A)$,*

$$\inf_{f \in \mathbb{e}_\mu^{-1}(\phi)} U(f, \pi) = \underline{U}(\pi).$$

This result conflicts with the intuition that since the sample mean $\mathbb{e}_\mu(f)$ is a good approximation of the population mean $\mathbb{e}(f)$, the mean explainer should be robust to sampling error. The problem is that these are different questions. When he sees an explanation ϕ, the object of interest for a DM trying to make robust decisions is not the set of DGPs that are approximated by (and thus have approximately the same population mean as) some DGP with population mean ϕ. Rather, the DM is concerned with the set of DGPs that are consistent with ϕ being the sample mean given some small sampling error. That is, robustness is over the set of DGPs $\bigcup_{\mu \in \mathcal{M}} \mathbb{e}_\mu^{-1}(\phi)$ for some neighborhood $\mathcal{M}$ of μ_0, *not* over the DGPs in some neighborhood of $\mathbb{e}^{-1}(\phi)$. Theorem 5 shows that this is an important distinction: even though $\mathbb{e}_\mu$ approaches $\mathbb{e}$ as μ approaches μ_0, the convergence is not uniform, and so the preimage $\mathbb{e}_\mu^{-1}(\mathbb{e}_\mu(f))$ does not collapse to $\mathbb{e}^{-1}(\mathbb{e}(f))$.

Intuitively, the robust usefulness of the population mean $\phi = \mathbb{e}(f)$ relies on every possible DGP $\hat{f} = (\hat{X}, \hat{Y})$ in its preimage $\mathbb{e}^{-1}(\phi)$ producing the same expected output $\mathbb{E}_{\omega \sim \mu_0}[\hat{Y}(\omega)]$. Or, put differently, it relies on every element of the mean explainer's kernel $\ker \mathbb{e} := \{\tilde{f} \in F \mid \mathbb{e}(\tilde{f}) = 0\}$ having an expected output of zero. In the appendix, we show that there are distributions μ arbitrarily close to μ_0 such that for any policy $\pi \in \Delta(A)$, $\ker \mathbb{e}_\mu$ contains DGPs $\tilde{f} = (\tilde{X}, \tilde{Y})$ such that $\pi \cdot \mathbb{E}_{\omega \sim \mu_0}[\tilde{Y}(\omega)]$ is arbitrarily low.

In fact, we show something stronger: Propositions 9 and 10 show that the same is true for the OLS explainer $\overline{\Gamma}_\mu$ computed under a sampling distribution μ, and thus that the additional information about the relationship between inputs and outputs provided by OLS is not enough to recover robust usefulness in the presence of sampling error. Moreover, this is true even when the decision maker can use the covariate vector to target her action: Theorem 6 shows that once OLS is computed under μ rather than μ_0, even targeted policies that lie in the space of regressors no longer have payoffs pinned down by the explanation.

Theorem 6 (Sampled OLS Is Useless in Targeted Problems)**.** *In a targeted utilitarian decision problem, the robust usefulness of OLS is not robust to sampling error: for any weak-**

open $\mathcal{M} \subseteq \Delta([0,1])$ *containing* μ_0*, there exists* $\mu \in \mathcal{M}$ *such that explanation with* $\overline{\Gamma}_\mu$ *is useless for robust decisions. That is, for every* $\phi \in \Phi$ *and every targeted policy* $\pi : \mathcal{X} \to \Delta(A)$*,*

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

Intuitively, the OLS explainer is robustly useful in targeted problems because the same distribution μ_0 is used in the OLS orthogonality condition and the utilitarian DM's payoff. Sampling error breaks that link. The appendix constructs sampling distributions μ arbitrarily close to μ_0 that put a small amount of mass on μ_0-null states. Those states are enough to hold the sampled OLS explanation fixed while moving the population payoff in any direction, and they can be chosen so that under μ_0 the covariates are constant with arbitrarily high probability. Once that happens, a targeted policy is evaluated almost entirely at a single covariate value x_0, so that in the limit, the targeted problem collapses back to an untargeted one.

Sample Selection on Observables and Weighted Least Squares

The negative results in this section arise from a mismatch between the sampling distribution μ and the payoff-relevant distribution μ_0 that is restricted in its magnitude (by requiring μ to be in a neighborhood of μ_0) but not in its form. As it turns out, if we restrict the set of possible DGPs to those for which this sampling mismatch is entirely explained by observables, we can restore robust usefulness.

To this end, consider a restricted class of data generating processes $\hat{F} \subset F$ and a distribution $\mu \in \Delta([0,1])$. We say that $(\hat{F}, \mu)$ satisfies *sample selection on observables* (Heckman 1979) with known weight $\rho : \mathcal{X} \to (0, \infty)$ if μ_0 is absolutely continuous with respect to μ and for every DGP $f = (X, Y) \in \hat{F}$,

$$\frac{d\mu_0}{d\mu}(\omega) = \rho(X(\omega)) \quad \text{for } \mu\text{-a.e. } \omega.$$

That is, the likelihood ratio between the population and sampling distributions depends only on the covariates. The same condition is known as *covariate shift* in the machine learning literature (Shimodaira 2000) and as *missing at random* in the statistics literature (Rubin 1976). Under this condition, the population objects from Sections 3.1 and 3.2 can be recovered exactly by reweighting the sample, as follows. Define the *reweighted mean explainer* as

$$\mathfrak{e}_\mu^\rho(f) := \mathbb{E}_{\omega \sim \mu}[\rho(X(\omega))Y(\omega)].$$

Given a space of regressors $\Phi \subset C(\mathcal{X}, \mathcal{Y})$, define the *weighted least squares (WLS) explainer* as

$$\overline{\Gamma}^{\rho}_{\mu}(f) \in \operatorname*{argmin}_{\phi \in \Phi} \mathbb{E}_{\omega \sim \mu} \left[\rho(X(\omega)) \|Y(\omega) - \phi(X(\omega))\|^2 \right],$$

and, for a basis b of the subspace $\hat{\Phi} \subset C(\mathcal{X})$, the *weighted Gram matrix* and the *augmented WLS explainer* $\widetilde{\overline{\Gamma}}^{b,\rho}_{\mu}$ by

$$\Sigma^{b,\rho}(f) := \mathbb{E}_{\omega \sim \mu} \left[\rho(X(\omega)) b(X(\omega)) b(X(\omega))^{\top} \right]; \qquad \widetilde{\Gamma}^{b,\rho}_{\mu}(f) = (\overline{\Gamma}^{\rho}_{\mu}(f), \Sigma^{b,\rho}(f)).$$

Proposition 6 (Reweighting Restores Population Objects)**.** *Suppose $(\hat{F}, \mu)$ satisfies sample selection on observables with known weight ρ. Then*

$$\mathbb{e}^{\rho}_{\mu}(f) = \mathbb{e}(f), \qquad \overline{\Gamma}^{\rho}_{\mu}(f) = \overline{\Gamma}(f), \qquad \Sigma^{b,\rho}(f) = \Sigma^{b}(f)$$

for every $f \in \hat{F}$.

Intuitively, once the likelihood ratio between μ_0 and μ is known and depends only on observables, reweighting converts each sample expectation back into the corresponding population expectation. Weighted least squares is therefore just population least squares written under a different measure.

Corollary 2 (Reweighted Means Are Robustly Perfect)**.** *In an untargeted utilitarian decision problem where the space of DGPs is restricted to $\hat{F}$, if $(\hat{F}, \mu)$ satisfies sample selection on observables with known weight ρ, then explanation with $\mathbb{e}^{\rho}_{\mu}$ is robustly perfect. In particular, the decision rule α^* from Theorem 1 always achieves the full-information benchmark.*

Corollary 3 (Weighted Least Squares Is Robustly Useful)**.** *In a targeted utilitarian decision problem where the space of DGPs is restricted to $\hat{F}$, if $(\hat{F}, \mu)$ satisfies sample selection on observables with known weight ρ, then explanation with the weighted least squares explainer $\overline{\Gamma}^{\rho}_{\mu}$ is robustly useful.*

Corollary 4 (Augmented WLS Is Asymptotically Robustly Perfect)**.** *For each $n \geq 1$, let $\widetilde{\Gamma}^{b,\rho}_{\mu,n}$ be an augmented WLS explainer with the space of explanations $\overline{\Phi}_n := \hat{\Phi}_n^M$, where $\hat{\Phi}_n$ is the space of polynomials on $\mathcal{X}$ of total degree at most n. In a targeted utilitarian decision problem where the space of DGPs is restricted to $\hat{F}$, if $(\hat{F}, \mu)$ satisfies sample selection on observables with known weight ρ, then the sequence $\{\widetilde{\Gamma}^{b,\rho}_{\mu,n}\}_{n \geq 1}$ is asymptotically robustly perfect.*

These corollaries restore the positive results from Sections 3.1 and 3.2 under sampling error, but only because of the structure that sample selection on observables imposes on the sample distribution. Without such structure, Theorems 5 and 6 show that even arbitrarily small sampling error can destroy robust usefulness altogether.

Example 1 revisited. A policymaker is deciding the eligibility criteria for a new government program, and observes an OLS regression of the program's potential outcomes on demographic characteristics x. If she only observes the OLS estimates from a sample distribution μ rather than the true population distribution μ_0, then even if the former is arbitrarily close to the latter, the worst-case total surplus from *any* policy could be arbitrarily bad. But if she knows that sampling error is determined entirely by observable characteristics, and she can recover the weighting function ρ, then weighted least squares estimates are robustly useful, and can be made even more useful by augmentation with the weighted gram matrix.

4 Ambiguity Aversion and Worst-Case Decision Problems

Theorem 1 shows that when the DM is utilitarian and the decision problem is untargeted, the mean explainer $\mathbb{e}$ achieves expected payoffs indistinguishable from those obtained with complete knowledge of the true DGP. Indeed, treating the mean explanation as if it were the true expected outcome vector yields the same expected payoff $\overline{U}(f)$ as having direct knowledge of the true DGP itself. And while the augmented OLS explainer $\widetilde{\Gamma}^b$ is not robustly perfect in targeted utilitarian problems, it is robustly useful — indeed, more useful for robust decisions than either the OLS or the mean explainer (Theorem 3). Furthermore, as more and more regressors are added to its space of explanations, $\widetilde{\Gamma}^b$ approaches robust perfection asymptotically (Theorem 4).

But data generating processes must often be explained to a decision maker who does not evaluate policies by their expected payoff under a single, known distribution. Policymakers may want to ensure that an intervention has beneficial effects on *all* members of a population, not just on average (i.e., they may have a *Rawlsian* social welfare function). Likewise, regulators or firms may be most concerned about the most catastrophic effects that could result from an AI model, not just the model's average performance. More generally, the decision maker may be *ambiguity-averse* in the sense of Gilboa and Schmeidler (1989): she does not know the probability measure μ_0 on the underlying probability space, holds a set of priors $\mathcal{M}$ instead, and wants to make a decision that is robust across DGPs *and also* across the distributions in $\mathcal{M}$.

Suppose that the true DGP is $f = (X, Y)$. Then in an untargeted decision problem with ambiguity aversion, a policy $\pi \in \Delta(A)$ yields indirect utility

$$U_{\mathcal{M}}(f, \pi) := \inf_{\mu \in \mathcal{M}} \mathbb{E}_{\omega \sim \mu}[Y(\omega) \cdot \pi],$$

whereas in a targeted problem with ambiguity aversion, a policy $\pi : \mathcal{X} \to \Delta(A)$ yields indirect

utility

$$U_{\mathcal{M}}(f,\pi) := \inf_{\mu\in\mathcal{M}} \mathbb{E}_{\omega\sim\mu}[\pi(X(\omega))\cdot Y(\omega)]\,.$$

A special case is *worst-case analysis*, in which the DM evaluates a policy by its worst payoff across all states of the world.

Example 8 (Rawlsian / Worst-Case Decision Problems)**.** Let the set of priors be the set of all point masses,

$$\underline{\mathcal{M}} := \{\delta_\omega : \omega \in [0,1]\},$$

where δ_ω denotes the Dirac measure at ω. Since the expectation of a random variable with respect to δ_ω simply evaluates it at ω, the ambiguity-averse indirect utility reduces to the worst payoff across states:

$$U_{\underline{\mathcal{M}}}(f,\pi) = \inf_{\omega\in[0,1]} \pi\cdot Y(\omega) =: R(f,\pi)$$

in an untargeted problem, and $U_{\underline{\mathcal{M}}}(f,\pi) = \inf_{\omega\in[0,1]} \pi(X(\omega))\cdot Y(\omega) =: R(f,\pi)$ in a targeted problem. We refer to R as *Rawlsian* indirect utility and write $\underline{R}(\pi) := \inf_{f\in F} R(f,\pi)$ for the corresponding no-information benchmark.

A Rawlsian's payoff reflects maximal ambiguity: The same indirect utility arises if the DM entertains every distribution on the underlying probability space, and so $\mathcal{M} = \Delta([0,1])$. Indeed, since $\mathbb{E}_{\omega\sim\mu}[\pi\cdot Y]$ is linear in μ, taking its infimum over $\Delta([0,1])$ gives the same value as taking the infimum over the set $\underline{\mathcal{M}}$ of extreme points of $\Delta([0,1])$. Thus $U_{\Delta([0,1])} = U_{\underline{\mathcal{M}}} = R$: worst-case analysis can equivalently be viewed as Rawlsian preferences over states or as complete ignorance about the prior.

Recall that an explainer's usefulness for robust decisions comes from its ability to produce explanations that identify policies. Unlike in the utilitarian case, the mean, OLS, and augmented OLS explainers do not accomplish this task in ambiguity averse decision problems with rich enough set of priors: in fact, no explanation from any of these explainers ever identifies any policy. The key reason is that each of these explainers is *affine* in Y: i.e., for any X, the map $Y \mapsto \Gamma(X,Y) - \Gamma(X,0)$ is linear. Theorem 7 shows that any such explainer is robustly useless in an ambiguity averse decision problem (targeted or untargeted) where the decision maker's set of priors $\mathcal{M}$ has dimension greater than the dimension of $\Gamma(F)$.

Theorem 7 (Affine Explainers are Not Robustly Useful with Ambiguity Aversion)**.** *In a decision problem with ambiguity aversion, if* Γ *is affine in* Y *and* $\dim(\mathcal{M}) \geq \dim(\Gamma(F))$*, then explanation with* Γ *is not useful for robust decisions. In particular, no explanation*

$\phi \in \Gamma(F)$ identifies any policy $\pi \in \Pi$:

$$\inf_{f \in \Gamma^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) = \underline{U}_{\mathcal{M}}(\pi) \text{ for all } \phi \in \Gamma(F) \text{ and } \pi \in \Pi.$$

In fact, the DM can do no better than naïvely maximizing her worst-case expected payoff over all covariates and outcomes:

$$\max_{\pi \in \Pi} \inf_{f \in \Gamma^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) = \max_{\pi \in \Pi} \underline{U}_{\mathcal{M}}(\pi) = \max_{\pi \in \Pi} \inf_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \pi(x) \cdot y. \tag{6}$$

Theorem 7 shows that explainers like $\mathfrak{e}$, $\overline{\Gamma}$, and $\widetilde{\Gamma}^b$ that are affine in Y can only be robustly useful to an ambiguity-averse DM if $\dim(\mathcal{M}) < \dim(\Gamma(F))$. Intuitively, an affine explainer discards information about at most $\dim(\Gamma(F))$ linearly independent functionals of the outcome process. The ambiguity-averse decision maker, on the other hand, cares about the set $\{\mathbb{E}_{\omega \sim \mu}[\pi(X(\omega)) \cdot Y(\omega)]\}_{\mu \in \mathcal{M}}$ — a family of functionals of Y that has $\dim(\mathcal{M}) + 1$ linearly independent elements.[4] Whenever $\dim(\mathcal{M}) \geq \dim(\Gamma(\mathcal{F}))$, at least one moment the decision maker cares about is not pinned down by the explanation she receives.

In particular, suppose the DM observes an explanation ϕ^* and chooses a policy π. Proposition 7 below shows that if $\dim(\mathcal{M}) \geq \dim(\Gamma(F))$, then for every possible value z, there is a prior $\mu \in \mathcal{M}$ and a DGP $f \in \Gamma^{-1}(\phi^*)$ consistent with the explanation under which π yields expected payoff z. Since z is arbitrary, it follows that the payoff guarantee $\inf_{f \in \Gamma^{-1}(\phi^*)} U_{\mathcal{M}}(f, \pi)$ can never exceed the no-information benchmark $\underline{U}_{\mathcal{M}}$.

Proposition 7. *Let $\Gamma : F \to \Phi$ be an explainer that is affine in Y; let $\phi^* \in \Gamma(F)$ be an explanation; let $q : \mathcal{X} \to \Delta(A)$ be a bounded Borel function; and let $z \in \mathbb{R}$. The set of priors*

$$\mathcal{M}_{z,q,\phi^*} := \{\mu \in \Delta([0,1]) \mid \mathbb{E}_{\omega \sim \mu}[q(X(\omega)) \cdot Y(\omega)] \neq z \ \forall (X, Y) \in \Gamma^{-1}(\phi^*)\} \tag{7}$$

has finite dimension less than $\dim(\Gamma(F))$.

Theorem 7 also provides a pessimistic perspective on Theorems 1 and 2 that complements that of Theorem 5. Each result shows, in slightly different ways, that Theorems 1 and 2 rely on the distribution used to compute the DM's payoff *exactly matching* the distribution used to compute the mean explanation. To show that robust explanation is not robust to sampling error, Theorem 5 perturbs the latter; to show that it is not robust to ambiguity about the distribution, Theorem 7 perturbs the former.

[4] Recall that the dimension of a set is the dimension of the smallest affine subspace containing it. Since $\mathcal{M}$ is a subset of the simplex $\Delta([0,1])$, it has dimension k when it contains $k+1$ linearly independent elements.

When applied to the worst-case (Rawlsian) problem described in Example 8, Theorem 7 yields a stronger corollary. Because the Dirac measures $\{\delta_\omega\}_{\omega\in[0,1]}$ are linearly independent, $\underline{\mathcal{M}}$ is infinite-dimensional, so the hypothesis $\dim(\underline{\mathcal{M}}) \geq \dim(\Gamma(F))$ holds automatically—no matter how rich the (finite-dimensional) space of explanations is. Theorem 7 therefore specializes immediately to the following.

Corollary 5 (Affine Explainers are Not Robustly Useful in Rawlsian Problems)**.** *In a Rawlsian decision problem, any explainer* Γ *that is affine in* Y *is useless for robust decisions. In particular, no explanation* $\phi \in \Gamma(F)$ *identifies any policy* $\pi \in \Pi$*:*

$$\inf_{f\in\Gamma^{-1}(\phi)} R(f,\pi) = \underline{R}(\pi) \text{ for all } \phi \in \Gamma(F) \text{ and } \pi \in \Pi.$$

In fact, the decision maker can do no better than naively maximizing her worst-case payoff over all covariates and outcomes:

$$\max_{\pi\in\Pi} \inf_{f\in\Gamma^{-1}(\phi)} R(f,\pi) = \max_{\pi\in\Pi} \underline{R}(\pi) = \max_{\pi\in\Pi} \inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x)\cdot y. \tag{8}$$

Unlike Theorem 7, the worst-case impossibility result of Corollary 5 does not weaken as the dimension of Φ increases. No matter how large the space of explanations is, *any* explanation from *any* explainer that is affine in Y provides no information useful for decisions that are robust across both states and DGPs — and this is true no matter how much the DM can observe about the covariates. For instance, if Φ is the set of nth degree polynomials, then no matter how large n is, no explainer improves the payoff of a DM who cares about the worst case, because the worst-case payoff consistent with an explanation is the same, no matter the explanation.

Remark (Garblings Inherit Uselessness). The reach of Theorem 7 and Corollary 5 extends beyond explainers that are affine in Y to those that can be derived from them. That is, if $\Gamma = h \circ \widetilde{\Gamma}$ for some map $h : \Phi \to \Psi$, and $\widetilde{\Gamma}$ is affine in Y, then the conclusions of Theorem 7 and Corollary 5 extend to Γ. As noted in Section 3, such a garbling (in the Blackwell sense) cannot be more useful for decision-making than the original explainer, because it must induce a coarser partition on the space of data generating processes. Since the worst-case payoff over each element of the partition induced by $\widetilde{\Gamma}$ is equal to the no-information benchmark, the same must be true of the worst-case payoff over each element of the coarser partition induced by Γ.

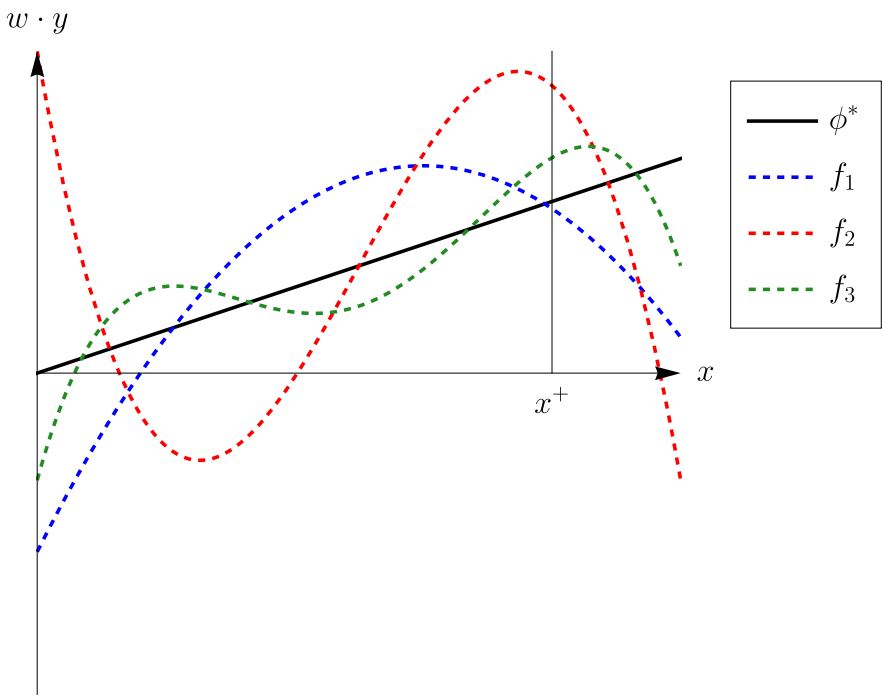

Figure 2: True models consistent with the same explanation. In the worst-case problem of Corollary 5, Proposition 7 implies that given an explainer Γ, for almost every state ω and any explanation ϕ^*, one can take any value z of the worst-case output from policy π and find a DGP consistent with ϕ^* such that $Y(\omega) \cdot \pi(X(\omega)) = z$. The figure illustrates an example where $\mathcal{X} = [0, 1]$, Γ is the OLS explainer $\overline{\Gamma}$, μ_0 is uniform, and $X(\omega) = \omega$. At $\omega = x^+$, the DGPs f_1, f_2, and f_3 induce very different values of $\pi \cdot y$, but each is consistent with the explanation ϕ^*: $\overline{\Gamma}(f_1) = \overline{\Gamma}(f_2) = \overline{\Gamma}(f_3) = \phi^*$.

Beyond Affine Explainers

The uselessness results of Theorem 7 and Corollary 5 both apply to explainers that are affine in outcomes, such as the mean or OLS explainers. This property is key to both results: An affine statistic such as the mean reports an *expected* outcome, whereas a worst-case or ambiguity-averse decision maker is concerned with a *lower bound* on her payoff, and the former does not offer any guarantees about the latter. Moving beyond affineness, on the other hand, can make explanation useful, even to an ambiguity-averse or Rawlsian decision maker. Indeed, if the decision maker wants to know the lower bound for the outcome of each action, an explainer can report that guarantee directly. More generally, an explainer can report a piecewise lower bound for the expected outcome conditional on a partition of the space of covariates. We call such lower-bound explanations *certificates*. Certificate explainers are not affine in Y — they are infima of expected payoffs — but they still map to a finite-dimensional space. And unlike any affine explainer, they can be robustly useful to

an ambiguity-averse decision maker.

The fundamental building blocks of these explainers are payoff *floors*. The simplest of these — and the ones that are relevant in an untargeted decision problem — are the *action floors*. Formally, for a set of priors $\mathcal{M}$ and each action $a \in A$, define

$$m_a^{\mathcal{M}}(f) := \inf_{\mu \in \mathcal{M}} \mathbb{E}_{\omega \sim \mu}[Y_a(\omega)],$$

the worst expected payoff across the priors in $\mathcal{M}$ from taking action a. In the Rawlsian special case $\mathcal{M} = \underline{\mathcal{M}}$, the floor is the worst outcome the action ever produces, $m_a^{\underline{\mathcal{M}}}(f) = \inf_{\omega \in [0,1]} Y_a(\omega)$. The *action-certificate explainer* reports the vector of floors, $\underline{\Gamma}_{A,\mathcal{M}}(f) := \left(m_a^{\mathcal{M}}(f)\right)_{a \in A} \in \mathbb{R}^A$. Like the mean explainer, $\underline{\Gamma}_{A,\mathcal{M}}$ is a statistic; unlike the mean explainer, it reports an infimum of expected outcomes rather than an expected outcome. This allows it to identify each untargeted policy, but not exactly.

Theorem 8 (Action Certificates Are Robustly Useful)**.** *In an untargeted decision problem with ambiguity aversion, the action-certificate explainer $\underline{\Gamma}_{A,\mathcal{M}}$ identifies each $\pi \in \Delta(A)$, with*

$$\inf_{f \in \underline{\Gamma}_{A,\mathcal{M}}^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) = \pi \cdot \phi.$$

Hence, explanation with $\underline{\Gamma}_{A,\mathcal{M}}$ is robustly useful: in particular, with the naïve decision rule $\alpha^*(\phi) \in \arg\max_{\pi \in \Delta(A)} \pi \cdot \phi$, for all $\pi \in \Delta(A)$ we have

$$\inf_{f \in \underline{\Gamma}_{A,\mathcal{M}}^{-1}(\phi)} U_{\mathcal{M}}(f, \alpha^*(\phi)) \geq \inf_{f \in \underline{\Gamma}_{A,\mathcal{M}}^{-1}(\phi)} U_{\mathcal{M}}(f, \pi),$$

with strict inequality for some $\phi \in \mathbb{R}^A$.

Each certificate ϕ identifies each policy π's worst-case payoff as $\pi \cdot \phi$: every DGP consistent with ϕ guarantees a payoff at least this high, because the floors bound each action's expected payoff from below, and the DGP whose outcome process is constant at ϕ guarantees exactly this payoff. Robust usefulness follows by comparing these guarantees across policies: the naïve rule $\alpha^*(\phi) \in \operatorname{argmax}_{\pi \in \Delta(A)} \pi \cdot \phi$ secures the largest of them, $\max_\pi \pi \cdot \phi$, which weakly exceeds the guarantee $\pi \cdot \phi$ of every other policy and strictly exceeds it for some ϕ.

We note that while the action certificate explainer is robustly useful, it is not robustly perfect. This is because reporting the floors of the pure actions can understate the payoff that can be guaranteed by mixing between them. For instance, suppose $\mathcal{M} = \{\mu_1, \mu_2\}$ with $\mathbb{E}_{\mu_1}[Y] = (1, 0)$ and $\mathbb{E}_{\mu_2}[Y] = (0, 1)$. Both action floors equal 0, so the certificate reports $(0, 0)$; yet the mixed policy $\pi = (\frac{1}{2}, \frac{1}{2})$ guarantees $\frac{1}{2}$ under both priors.

Targeted problems. When the decision maker can condition her action on the covariates, the analogue of the action floor is a conditional floor on a partition of the space $\mathcal{X}$ of covariates. Fix a finite partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ of $\mathcal{X}$ such that each C_k is a Borel set with nonempty interior. For each action a and cell C, define the *conditional floor*

$$\ell^{\mathcal{M}}_{a,C}(X,Y) := \inf_{\substack{\mu\in\mathcal{M}:\\ \mu(X^{-1}(C))>0}} \mathbb{E}_{\omega\sim\mu}\big[Y_a(\omega) \mid X(\omega)\in C\big],$$

the worst expected payoff across the priors in $\mathcal{M}$, conditional on the covariates being in C, from taking action a.[5] The *certificate explainer* reports the map from covariates to outcomes that is generated piecewise by the conditional floors:

$$\underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(f) := \big(\sum_{C\in\mathcal{C}} \ell^{\mathcal{M}}_{a,C}(f)\mathbf{1}_C\big)_{a\in A} \in \left(\operatorname{span}\{\mathbf{1}_C\}_{C\in\mathcal{C}}\right)^M.$$

In the Rawlsian special case $\mathcal{M} = \underline{\mathcal{M}}$, each floor is simply the worst outcome the action ever produces on its cell, $\ell^{\underline{\mathcal{M}}}_{a,C}(f) = \inf_{\omega\,:\,X(\omega)\in C} Y_a(\omega)$. Like OLS, the certificate is a model that provides information about the expectation of Y conditional on X; unlike OLS, it provides a lower bound for the conditional expectation, rather than an approximation to it.

As Theorem 9 shows, $\underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}$ is robustly useful because it allows the DM to use a decision rule $\alpha^{\mathcal{C}}$ which chooses a policy that is constant on each cell C at the action a with the highest certificate value $\phi_{a,C}$.[6]

Theorem 9 (Partition Certificates Are Robustly Useful)**.** *In a targeted decision problem with ambiguity aversion, the partition-certificate explainer $\underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}$ is robustly useful. In particular, for every targeted policy $\pi : \mathcal{X} \to \Delta(A)$ and every explanation $\phi \in \underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(F)$,*

$$\inf_{f\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f,\alpha^{\mathcal{C}}(\phi)) \geq \inf_{f\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f,\pi),$$

and, for every π, this inequality is strict for some $\phi \in \underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(F)$.

The argument for Theorem 9 is the cell-by-cell counterpart of Theorem 8. On any cell with finite certificate values, every DGP consistent with a partition certificate ϕ yields an expected payoff of at least $\phi_{a,C}$ from action a. Thus, for any prior and any DGP in the explanation cell, the rule $\alpha^{\mathcal{C}}$ yields at least the prior-weighted average of the cellwise maxima $\max_a \phi_{a,C}$ over the cells that receive positive probability. Conversely, holding fixed the covariate process and setting $Y \equiv \phi_C$ on each relevant cell gives an upper bound on the robust payoff guarantee

[5] Recall that $\inf\emptyset = +\infty$.

[6] Formally, $\alpha^{\mathcal{C}}(\phi) := \pi^{\phi}$, where $\pi^{\phi}(x) \in \operatorname{argmax}_{a\in A}\phi_{a,C}$ for each $x \in C$.

of any alternative policy. Maximizing $\pi(C) \cdot \phi_C$ separately in each cell thus yields the best robust guarantee that the certificate can provide.

Partition certificates are robustly useful but, like OLS in the utilitarian problem, not robustly perfect. First, outcomes can vary within any given cell. Second, nature chooses a single prior globally while the certificate minimizes separately across actions and cells. Refining the partition $\mathcal{C}$ can shrink the first gap, but for the same reason that the action-certificate explainer is not robustly perfect in untargeted problems, the second gap may remain.

Example 2 revisited. A regulator is concerned about the cybersecurity risks of allowing a new AI model to be deployed. It is not focused on the risks on average, but on the worst possible risks that it considers to be within the realm of plausibility. That is, it is ambiguity averse: it behaves as if it is facing an adversary that can shift probability mass within $\mathcal{M} \subseteq \Delta([0,1])$ so that the data generating process f generates prompts that are less likely to trigger guardrails and responses that are more likely to be harmful.

Whether it can regulate these guardrails at the prompt level (a targeted policy), or can only issue regulation at the model level (an untargeted policy), a statistic describing average harm — or any method of explaining the model's behavior that is affine in the level of harm — is not robustly useful to such a regulator. But a guarantee of the model's behavior against an adversary, on the other hand, is robustly useful to the ambiguity averse or Rawlsian regulator.

5 Discussion

Our main results are summarized in Table I. For a decision maker who maximizes average payoffs, explanations that are affine in outcomes, namely the mean of the outcome process and the least squares regression model, are useful for robust decisions. For a DM who maximizes worst-case payoffs, by contrast, every affine explanation is useless for rich enough priors, while certificates, which report guarantees rather than averages, remain useful for robust decisions.

These results have implications for the interpretation of empirical work. Many empirical analyses aim to be as agnostic as possible about the way that the variables that they study are related. Instead of making *structural* assumptions about the form of that relationship, they provide a *reduced form* estimate of a linear approximation to it. For instance, if analyses wish to estimate the effect of a treatment à la Example 1, they might not make assumptions about the joint distribution of the treatment effect and individual characteristics (i.e., limitations on the space of true models F, or a prior over F) and instead focus on estimating an average

	Untargeted	Targeted
Utilitarian	Mean: robustly perfect (Theorem 1).	OLS: robustly useful (Theorem 2); augmented OLS: asymptotically robustly perfect (Theorem 4).
Worst-case (ambiguity averse)	Any affine explainer: useless for rich enough priors (Theorem 7); certificates: robustly useful (Theorem 8).	Any affine explainer: useless for rich enough priors (Theorem 7); certificates: robustly useful (Theorem 9).

Table I: Robustly useful explanations, by payoff criterion and policy type. For a decision maker who maximizes average payoffs, explanations affine in outcomes (the mean and least squares) are useful for robust decisions; for one who maximizes worst-case payoffs, affine explanations are useless for rich enough priors while certificates that report guarantees rather than averages remain useful.

treatment effect (i.e., the expectation $\mathbb{E}_{\omega\sim\mu_0}[Y(\omega)]$) or conditional average treatment effect (i.e., the conditional expectation $\mathbb{E}_{\omega\sim\mu_0}[Y(\omega)|X(\omega) = x]$). Similarly, they might appeal to the central limit theorem to avoid making assumptions about the distribution of sampling error.

Our results give a sharp characterization of the usefulness of this approach to decision makers who rely on the results of such analyses. When there is no sampling error, and decision makers focus on the *average* effects of their decisions, Theorems 1-3 show that structural assumptions are not required for robust usefulness. But when sampling error is present, Theorem 5 shows that this is not necessarily true: Sample averages and least-squares estimates can only be useful for decision making when combined with assumptions about the data generating process and/or the form of the sampling error. Likewise, when a decision maker is interested in a worst-case rather than average effect, Corollary 5 shows that even in the absence of sampling error, least squares estimates are only useful when combined with structural assumptions.

However, Corollaries 2-4 show how a theoretical assumption about the data generating process can be combined with sample data to produce reports that are robustly useful to a decision maker. In particular, if sample selection is completely explained by the covariates X observable to the DM — a *joint* assumption on the space of DGPs and the sampling error $\frac{d\mu_0}{d\mu}$ — then when they are computed under a suitable reweighting, least-squares estimates and sample means are once again useful for robust decisions. Hence, even when a statistician wants to avoid making structural assumptions about relationships between variables, theo-

retical assumptions about relationships between the covariates in the data and selection into the sample may still be very useful.

Conversely, our positive results in Section 4 offer a theory-agnostic way of offering robust explanations to an ambiguity-averse decision maker or a Rawlsian who cares about worst-case effects. Instead of using a parametric model or a moment of the outcome distribution, such decision-makers benefit from nonparametric *certificates* that offer a lower bound on the payoff from each of their actions, either in general (Theorem 8) or conditional on observable covariates (Theorem 9).

6 Related Literature

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

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

Our targeted results relate to statistical treatment rules and empirical welfare maximization (Manski 2004; Hirano and Porter 2009; Stoye 2009, 2012; Tetenov 2012; Kitagawa and Tetenov 2018; Athey and Wager 2021). That literature starts from sample data, studying how a planner should map data into a treatment rule, typically using welfare or regret as the criterion. Kitagawa and Tetenov (2018), for example, maximize sample welfare over a fixed class of policies, and Athey and Wager (2021) study policy learning from observational

data with constraints for budget, fairness, or simplicity. We instead fix the decision problem and ask which explanations support robust choice. The worst case is over the DGPs consistent with an explanation, not over sampling distributions. And the class $\Pi_{\overline{\Phi}}$ is not imposed externally—it is the set of targeted policies whose payoffs are identified by an OLS explanation.

Finally, there is a large literature in computer science on the use of explanations to help humans use predictions made by AI models to make decisions (e.g., Lai and Tan 2019; Bansal, Wu, Zhou, Fok, Nushi, Kamar, Ribeiro and Weld 2021). This focus is subtly different than ours: While we also consider the explanation of the joint distribution of covariates (e.g., features) and outcomes generated by an AI model (e.g., Example 2), in our setting this data generating process is directly relevant to the DM's payoff, rather than providing suggestions that inform the DM about another DGP. The explanations in our paper thus provide information about the underlying payoff-relevant relationship directly, instead of indirectly (by providing information about an approximating model).

7 Conclusion

We consider the problem of explaining data generating processes to a decision maker (DM). The DM's payoff depends on her action and the state of the world, where the latter is described by covariates and outcomes. A data generating process (DGP) specifies the joint distribution of these covariates and outcomes, but is not intelligible to the DM. For the DM to make a choice, the DGP instead has to be *explained* to her using a report that belongs to a finite-dimensional space. We show that if the DM maximizes her average payoff, then the mean outcome is as good as understanding the DGP itself when actions cannot be targeted to the covariates; when actions can be targeted, least-squares regressions augmented with the second moments of their regressors come arbitrarily close as the space of regressors grows rich. These guarantees can fail under arbitrarily small sampling error, but they are restored when selection into the sample depends on the covariates in a known way. If the DM instead maximizes her worst-case payoff, then any explanation that is affine in outcomes — including the mean and least squares — offers no advantage over no explanation at all. Explanations that report payoff guarantees, which we call certificates, are useful even to such a decision maker.

The paper's environment leaves room for continuing work. For example, we focus on a single decision maker, but a second agent could be introduced, one who provides explanations of DGPs that may misalign with the interests of the decision maker.[7] Second, it remains

[7]Recently, Liang, Lu and Mu (2023) examine algorithmic fairness in an information design setting, where

to be seen whether other assumptions on the space of DGPs or the form of sampling error, besides the sample selection on observables explored in Section 3.3, allow explanations to remain useful for robust decisions.

Finally, one could study explainers beyond those considered here. Neural networks are a natural candidate: a network trained to approximate a real-world DGP is a finite-dimensional report of that DGP, just as a fitted regression is, and the same is true of a small network trained to approximate a larger one, as in model distillation. Unlike the mean and least squares, the map from the DGP to the fitted network is not affine in outcomes, so our impossibility results with ambiguity aversion (Theorem 7 and Corollary 5) do not apply to it. Instead, which policies such an explanation identifies, and for which decision makers it is useful, are open questions.

We could also consider Monte Carlo-based explainers in reinforcement learning. Instead of identifying the payoffs of targeted policies in a fixed class (as we show OLS and augmented OLS do in Theorem 3), Monte Carlo methods such as the structural reinforcement learning of Yang, Wang, Schaab and Moll (2026) instead report the payoff of a parameterized policy and the gradient of that payoff with respect to the parameters, and improve the policy by gradient ascent. In our framework, such a method identifies the payoffs of precisely the policies encountered during the ascent process — a class that is determined endogenously by the algorithm, rather than fixed in advance.

References

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

Anthropic PBC (2026) "System Card: Claude Fable 5 & Claude Mythos 5,"Technical report, https://www-cdn.anthropic.com/2f9323abbcc4abe219577539efe19a623c9ca2bd/Claude%20Fable%205%20&%20Claude%20Mythos%205%20System%20Card.pdf, Accessed: June 19, 2026.

Athey, S. and S. Wager (2021) "Policy Learning with Observational Data," *Econometrica*, 89 (1), 133–161.

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.

a sender chooses inputs to an algorithm and a receiver chooses the algorithm.

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

Blackwell, D. (1953) "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 24 (2), 265–272, http://www.jstor.org/stable/2236332.

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

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

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

Cheng, X. (2026) "Improving Robust Decisions with Data," Forthcoming, *Theoretical Economics*.

Dugundji, J. (1951) "An Extension of Tietze's Theorem," *Pacific Journal of Mathematics*, 1 (3), 353–367.

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

Fudenberg, D., W. Gao, and A. Liang (2026) "How flexible is that functional form? Quantifying the restrictiveness of theories," *The Review of Economics and Statistics*, 108 (1), 194–209.

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.

Heckman, J. J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1), 153–161.

Hirano, K. and J. R. Porter (2009) "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77 (5), 1683–1701.

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

Kitagawa, T. and A. Tetenov (2018) "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86 (2), 591–616.

Lai, V. and C. Tan (2019) "On Human Predictions With Explanations and Predictions of Machine

Learning Models: A Case Study On Deception Detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

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

Lorentz, G. G. (1986) *Bernstein Polynomials*: Chelsea Publishing Company, 2nd edition.

Manski, C. F. (2004) "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72 (4), 1221–1246.

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

Rubin, D. B. (1976) "Inference and Missing Data," *Biometrika*, 63 (3), 581–592.

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.

Shimodaira, H. (2000) "Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function," *Journal of Statistical Planning and Inference*, 90 (2), 227–244.

Stoye, J. (2009) "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151 (1), 70–81.

——— (2012) "Minimax Regret Treatment Choice With Covariates or With Limited Validity of Experiments," *Journal of Econometrics*, 166 (1), 138–156.

Tetenov, A. (2012) "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," *Journal of Econometrics*, 166 (1), 157–165.

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

Yang, Y., C. Wang, A. Schaab, and B. Moll (2026) "Structural Reinforcement Learning for Heterogeneous Agent Macroeconomics," Working paper, https://benjaminmoll.com/wp-content/uploads/2025/12/SRL.pdf.

Proofs

Proof of Theorem 1 (Mean Explanations Are Robustly Perfect) Fix $f = (X, Y) \in F$. By definition of the mean explainer, for each $\pi \in \Delta(A)$,

$$\mathbb{e}(f) \cdot \pi = \mathbb{E}_{\omega \sim \mu_0}[Y(\omega)] \cdot \pi = U(f, \pi).$$

Therefore,

$$U(f, \alpha^*(\mathbb{e}(f))) = \mathbb{e}(f) \cdot \alpha^*(\mathbb{e}(f)) = \max_{\pi \in \Delta(A)} \mathbb{e}(f) \cdot \pi = \max_{\pi \in \Delta(A)} U(f, \pi) = \overline{U}(f).$$

Hence the decision rule α^* attains the full-information benchmark for every $f \in F$, so explanation with $\mathbb{e}$ is robustly perfect. ■

Lemma 2. *Let $\mathcal{S}$ be a set, and let $\{g_i : \mathcal{S} \to \mathbb{R}\}_{i=1}^N$ be linearly independent. There exist $\{x_i\}_{i=1}^N \subset \mathcal{S}$ such that the matrix $G_N(\{x_i\}_{i=1}^N) = \begin{bmatrix} g_1(x_1) & \cdots & g_N(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_N) & \cdots & g_N(x_N) \end{bmatrix}$ has full rank.*

Proof. We proceed by induction on N.

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

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

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

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

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

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

Lemma 3. *Let $x^-, x^+ \in \mathcal{X}$. Then there exists a continuous surjection $q : [0,1] \to \mathcal{X}$ such that $q(0) = x^-$ and $q(1) = x^+$.*

Proof. Because $\mathcal{X}$ is a compact convex subset of $\mathbb{R}^K$, it is compact, connected, and locally connected. By the Hahn–Mazurkiewicz theorem, there exists a continuous surjection $\widetilde{q} : [0,1] \to \mathcal{X}$. Let $u := \widetilde{q}(0)$ and $v := \widetilde{q}(1)$. Since $\mathcal{X}$ is convex, the line segments from x^- to u and from v to x^+ lie in $\mathcal{X}$. Define $q : [0,1] \to \mathcal{X}$ by

$$q(t) := \begin{cases} (1-3t)x^- + 3tu, & t \in [0, \frac{1}{3}], \\ \widetilde{q}(3t-1), & t \in [\frac{1}{3}, \frac{2}{3}], \\ (3-3t)v + (3t-2)x^+, & t \in [\frac{2}{3}, 1]. \end{cases}$$

The three expressions agree at $t = \frac{1}{3}$ and $t = \frac{2}{3}$, so q is continuous. Moreover, $q([0,1]) \supseteq \widetilde{q}([0,1]) = \mathcal{X}$, while $q([0,1]) \subseteq \mathcal{X}$ since $u, v, x^-, x^+ \in \mathcal{X}$ and $\mathcal{X}$ is convex. Thus $q([0,1]) = \mathcal{X}$, with $q(0) = x^-$ and $q(1) = x^+$. ■

Lemma 4. *Fix a continuous onto input process $X^0 : [0,1] \to \mathcal{X}$, let $\mu_{X^0} = X^0_{\#}\mu_0$, and let $\pi : \mathcal{X} \to \Delta(A)$ be a targeted policy that is not μ_{X^0}-almost surely equal to any policy in $\Pi_{\overline{\Phi}}$. Then for every $\phi \in \hat{\Phi}^M$ and every $z \in \mathbb{R}$, there exists a DGP $f^z = (X^0, Y^z)$ such that $\overline{\Gamma}(f^z) = \phi$ and $U(f^z, \pi) = z$.*

Proof. Because π is not μ_{X^0}-almost surely equal to any policy in $\Pi_{\overline{\Phi}}$, there is an action a such that π_a is not μ_{X^0}-almost surely equal to any element of $\hat{\Phi}$. Since π_a is bounded and Borel-measurable, $\pi_a \in L^2(\mu_{X^0})$. Let $P^{\mu_{X^0}}_{\hat{\Phi}}$ be the orthogonal projection of $L^2(\mu_{X^0})$ onto $\hat{\Phi}$, and let $r := \pi_a - P^{\mu_{X^0}}_{\hat{\Phi}}(\pi_a)$ be the residual from the projection of π_a onto $\hat{\Phi}$. Then r is bounded and Borel measurable, nonzero, orthogonal to $\hat{\Phi}$ in $L^2(\mu_{X^0})$, and $\int \pi_a r \, \mathrm{d}\mu_{X^0} = \int r^2 \, \mathrm{d}\mu_{X^0} > 0$. For $z \in \mathbb{R}$, define

$$Y^z(\omega) := \phi(X^0(\omega)) + \left(\frac{z - \int \pi \cdot \phi \, \mathrm{d}\mu_{X^0}}{\int r^2 \, \mathrm{d}\mu_{X^0}} \right) r(X^0(\omega)) e_a.$$

Since r and ϕ are bounded and Borel-measurable, Y^z is bounded and Borel-measurable for each z. Since r is orthogonal to $\hat{\Phi}$ in $L^2(\mu^0)$, we have

$$\int_{[0,1]} r(X^0(\omega)) \psi(X^0(\omega)) \, \mathrm{d}\mu_0(\omega) = 0 \text{ for all } \psi \in \hat{\Phi}.$$

It follows that $e_a \times r \circ X^0$ is orthogonal to Ψ_{X^0} in $L^2([0,1], \mu_0; \mathcal{Y})$, and so $\overline{\Gamma}(X^0, Y^z) = \phi$. Moreover,

$$U((X^0, Y^z), \pi) = \mathbb{E}_{\omega \sim \mu_0}[\pi(X^0(\omega)) \cdot \phi(X^0(\omega))] + \left(\frac{z - \int \pi \cdot \phi \, \mathrm{d}\mu_{X^0}}{\int r^2 \, \mathrm{d}\mu_{X^0}} \right) \int \pi_a r \, \mathrm{d}\mu_{X^0} = z,$$

as desired. ■

Lemma 5. *If $\pi \notin \Pi_{\overline{\Phi}}$, then there exists a continuous onto input process $X^0 : [0,1] \to \mathcal{X}$ such that π is not μ_{X^0}-almost surely equal to any policy in $\Pi_{\overline{\Phi}}$, where $\mu_{X^0} = (X^0)_{\#}\mu_0$.*

Proof. Let $a \in A$ be such that $\pi_a \notin \hat{\Phi}$, and let $\{\psi_j\}_{j=1}^{\dim(\hat{\Phi})}$ be a basis for $\hat{\Phi}$. By Lemma 2, there exist points $\{x_i\}_{i=1}^{\dim(\hat{\Phi})+1} \subset \mathcal{X}$ such that the matrix

$$\begin{bmatrix} \psi_1(x_1) & \cdots & \psi_{\dim(\hat{\Phi})}(x_1) & \pi_a(x_1) \\ \vdots & \ddots & \vdots & \vdots \\ \psi_1(x_{\dim(\hat{\Phi})+1}) & \cdots & \psi_{\dim(\hat{\Phi})}(x_{\dim(\hat{\Phi})+1}) & \pi_a(x_{\dim(\hat{\Phi})+1}) \end{bmatrix}$$

has full rank. Choose numbers $\{a_i, b_i\}_{i=1}^{\dim(\hat{\Phi})+1}$ such that

$$0 = a_1 < b_1 < a_2 < b_2 < \cdots < a_{\dim(\hat{\Phi})+1} < b_{\dim(\hat{\Phi})+1} = 1.$$

Let $I_i := [a_i, b_i]$ and $J_i := [b_i, a_{i+1}]$. Since μ_0 has full support, $\mu_0(I_i) > 0$ for every i. For each $i = 1, \ldots, \dim(\hat{\Phi})$, use Lemma 3 to choose a continuous surjection $q_i : [0,1] \to \mathcal{X}$ such that $q_i(0) = x_i$ and $q_i(1) = x_{i+1}$. Define X^0 by setting $X^0(\omega) = x_i$ on I_i and $X^0(\omega) = q_i((\omega - b_i)/(a_{i+1} - b_i))$ on J_i. Then X^0 is continuous and onto.

If π were μ_{X^0}-almost surely equal to some $\hat{\pi} \in \Pi_{\overline{\Phi}}$, then $\pi_a(x_i) = \hat{\pi}_a(x_i)$ for every i, because $\mu_{X^0}(\{x_i\}) \geq \mu_0(I_i) > 0$. This would make the last column of the displayed matrix a linear combination of the first $\dim(\hat{\Phi})$ columns, a contradiction. ■

Proof of Proposition 1 Fix $f \in \overline{\Gamma}^{-1}(\phi)$. Since $\pi \in \Pi_{\overline{\Phi}} = \overline{\Phi} \cap \Delta(A)^{\mathcal{X}}$, we have $\pi \circ X \in \Psi_X$. The OLS residual $Y - \phi \circ X$ is orthogonal to Ψ_X, so

$$U(f, \pi) = \mathbb{E}_{\omega \sim \mu_0} \left[\pi(X(\omega)) \cdot \phi(X(\omega)) \right] \geq \inf_{x \in \mathcal{X}} \pi(x) \cdot \phi(x).$$

Taking the infimum over $f \in \overline{\Gamma}^{-1}(\phi)$ yields

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi)} U(f, \pi) \geq \inf_{x \in \mathcal{X}} \pi(x) \cdot \phi(x).$$

Choose $x_0 \in \arg\min_{x \in \mathcal{X}} \pi(x) \cdot \phi(x)$, which is nonempty since $\pi \in \Pi_{\overline{\Phi}} \subset \Phi$ and $\phi \in \Phi$ are continuous and $\mathcal{X}$ is compact. Fix $\varepsilon > 0$. Since every probability measure on $[0,1]$ has at most countably many atoms, we can choose $\hat{\omega} \in (0,1)$ such that $\mu_0(\{\hat{\omega}\}) = 0$. By continuity from above, there exists a closed interval $I_\varepsilon = [a_\varepsilon, b_\varepsilon] \subset (0,1)$ containing $\hat{\omega}$ such that $\mu_0(I_\varepsilon) < \varepsilon$. By Lemma 3, there exists a continuous surjection $q_\varepsilon : [0,1] \to \mathcal{X}$ with $q_\varepsilon(0) = q_\varepsilon(1) = x_0$. Define a continuous onto input process $\hat{X}^\varepsilon : [0,1] \to \mathcal{X}$ by

$$\hat{X}^\varepsilon(\omega) := \begin{cases} x_0 & \text{if } \omega \notin I_\varepsilon, \\ q_\varepsilon\left(\frac{\omega - a_\varepsilon}{b_\varepsilon - a_\varepsilon}\right) & \text{if } \omega \in I_\varepsilon. \end{cases}$$

Let $\hat{Y}^\varepsilon := \phi \circ \hat{X}^\varepsilon \in \Psi_{\hat{X}^\varepsilon}$, and let $\hat{f}^\varepsilon := (\hat{X}^\varepsilon, \hat{Y}^\varepsilon)$. Then the projection of $\hat{Y}^\varepsilon$ onto $\Psi_{\hat{X}^\varepsilon}$ is just $\hat{Y}^\varepsilon = \phi \circ \hat{X}^\varepsilon$. Therefore $\hat{f}^\varepsilon \in \overline{\Gamma}^{-1}(\phi)$.

Since $\pi \in \Pi_{\overline{\Phi}} \subset \Phi$ and $\phi \in \Phi$ are continuous and $\mathcal{X}$ is compact, we have $\sup_{x \in \mathcal{X}} \pi(x) \cdot$

$\phi(x) < \infty$. Then

$$U(\hat{f}^\varepsilon, \pi) = \int_{[0,1]\setminus I_\varepsilon} \pi(x_0) \cdot \phi(x_0)\, \mathrm{d}\mu_0(\omega) + \int_{I_\varepsilon} \pi(\hat{X}^\varepsilon(\omega)) \cdot \phi(\hat{X}^\varepsilon(\omega))\, \mathrm{d}\mu_0(\omega)$$
$$\leq (1-\varepsilon) \inf_{x\in\mathcal{X}} \pi(x) \cdot \phi(x) + \varepsilon \sup_{x\in\mathcal{X}} \pi(x) \cdot \phi(x).$$

Hence, for every $\varepsilon > 0$,

$$\inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \pi) \leq \inf_{x\in\mathcal{X}} \pi(x) \cdot \phi(x) + \varepsilon \sup_{x\in\mathcal{X}} \pi(x) \cdot \phi(x).$$

The claim follows by letting $\varepsilon \downarrow 0$.

Proof of Proposition 2 By Lemma 5, there exists a continuous onto input process X^0 such that π is not μ_{X^0}-almost surely equal to any policy in $\Pi_{\overline{\Phi}}$, where $\mu_{X^0} = (X^0)_{\#}\mu_0$. By Lemma 4, for each $z \in \mathbb{R}$, there exists $f^z = (X^0, Y^z)$ such that $\overline{\Gamma}(f^z) = \phi$ and $U(f^z, \pi) = z$. Then for each $z \in \mathbb{R}$,

$$\inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \pi) \leq z.$$

Since z is arbitrary, it follows that the infimum on the left is $-\infty$. Since $\overline{\Gamma}^{-1}(\phi) \subseteq F$, it follows that $\underline{U}(\pi) = -\infty$ as well. ■

Proof of Theorem 2 (OLS Is Robustly Useful) Fix $\phi \in \Phi$. For each identified policy $\pi \in \Pi_{\overline{\Phi}}$, Proposition 1 gives

$$\inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \pi) = \inf_{x\in\mathcal{X}} \pi(x) \cdot \phi(x).$$

Hence, by the definition of α^{OLS},

$$\inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \alpha^{\mathrm{OLS}}(\phi)) \geq \inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \pi) \quad \text{for every } \pi \in \Pi_{\overline{\Phi}}.$$

If instead $\pi \notin \Pi_{\overline{\Phi}}$, Proposition 2 yields

$$\inf_{f\in\overline{\Gamma}^{-1}(\phi)} U(f, \pi) = \underline{U}(\pi) = -\infty,$$

so the same weak inequality holds for every targeted policy $\pi : \mathcal{X} \to \Delta(A)$.

It remains to verify strictness. Fix a targeted policy π. If $\pi \notin \Pi_{\overline{\Phi}}$, choose any action $a \in A$, and let $\phi^a(x) \equiv e_a$. The constant pure policy $x \mapsto e_a$ belongs to $\Pi_{\overline{\Phi}}$, and it is the

unique solution to $\max_{\pi' \in \Pi_{\overline{\Phi}}} \inf_{x \in \mathcal{X}} \pi'(x) \cdot \phi^a(x)$. Therefore

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi^a)} U(f, \alpha^{\text{OLS}}(\phi^a)) = 1 > -\infty = \inf_{f \in \overline{\Gamma}^{-1}(\phi^a)} U(f, \pi).$$

Suppose now that $\pi \in \Pi_{\overline{\Phi}}$. If π is not a constant pure policy, then there exists $a \in A$ such that $\inf_{x \in \mathcal{X}} \pi_a(x) < 1$. For $\phi^a(x) \equiv e_a$, Proposition 1 implies

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi^a)} U(f, \alpha^{\text{OLS}}(\phi^a)) = 1 > \inf_{x \in \mathcal{X}} \pi_a(x) = \inf_{f \in \overline{\Gamma}^{-1}(\phi^a)} U(f, \pi).$$

Finally, if π is the constant pure policy $x \mapsto e_a$, choose any $b \neq a$ and let $\phi^b(x) \equiv e_b$. Then

$$\inf_{f \in \overline{\Gamma}^{-1}(\phi^b)} U(f, \alpha^{\text{OLS}}(\phi^b)) = 1 > 0 = \inf_{f \in \overline{\Gamma}^{-1}(\phi^b)} U(f, \pi).$$

Thus explanation with $\overline{\Gamma}$ is useful for robust decisions. ■

Lemma 6. *In a utilitarian decision problem, $\underline{U}(\pi) = -\infty$ for each $\pi \in \Pi$.*

Proof. Fix $\pi \in \Pi$ and $z \in \mathbb{R}$. Let $Y(\omega)_a = z$ for each $a \in A$ and $\omega \in [0,1]$, and let $X : [0,1] \to \mathcal{X}$ be continuous and onto. Then whether the decision problem is targeted or untargeted, we have $U((\mu_0, Y), \pi) = z$. Since z is arbitrary, $\underline{U}(\pi) = -\infty$. ■

Proof of Proposition 3 Let $f \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma)$. Since $\pi = Gb \in \Pi_{\overline{\Phi}}$, $\pi \circ X \in \Psi_f$. The OLS error $Y - Bb \circ X$ is therefore orthogonal to $\pi \circ X$, and so

$$\begin{aligned} U(f, \pi) &= \mathbb{E}_{\omega \sim \mu_0} \left[(Gb(X(\omega)))^\top Bb(X(\omega)) \right] = \sum_{a \in A} \mathbb{E}_{\omega \sim \mu_0} \left[(G_{a\cdot} b(X(\omega)))(B_{a\cdot} b(X(\omega))) \right] \\ &= \sum_{a \in A} G_{a\cdot} \mathbb{E}_{\omega \sim \mu_0} \left[b(X(\omega)) b(X(\omega))^\top \right] (B_{a\cdot})^\top = \sum_{a \in A} G_{a\cdot} \Sigma (B_{a\cdot})^\top = \operatorname{tr}(G \Sigma B^\top). \end{aligned}$$

Since this is finite, Lemma 6 implies $U(f, \pi) = U(g, \pi) > \underline{U}(\pi) = -\infty$ for each $f, g \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma)$, as desired.

Proposition 8. *Let $(Bb, \Sigma) \in \hat{\Phi}^M \times \mathbb{R}^{\dim(\hat{\Phi}) \times \dim(\hat{\Phi})}$ be an augmented OLS explanation, and let $\pi : \mathcal{X} \to \Delta(A)$ be a targeted policy. If there exists a continuous onto input process $X^0 : [0,1] \to \mathcal{X}$ such that $\mathbb{E}_{\omega \sim \mu_0}[b(X^0(\omega)) b(X^0(\omega))^\intercal] = \Sigma$ and π is not μ_{X^0}-almost surely equal to any policy in $\Pi_{\overline{\Phi}}$, where $\mu_{X^0} = (X^0)_{\#}\mu_0$, then*

$$\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma)} U(f, \pi) = \underline{U}(\pi) = -\infty.$$

Proof. For every $z \in \mathbb{R}$, Lemma 4 gives a DGP $f^z = (X^0, Y^z)$ such that $\overline{\Gamma}(f^z) = Bb$ and $U(f^z, \pi) = z$; by assumption, $\Sigma^b(f^z) = \Sigma$. Thus $f^z \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma)$, so

$$\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f, \pi) \leq z \quad \text{for every } z \in \mathbb{R}.$$

Hence the left-hand side is $-\infty$. Since $(\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma) \subseteq F$, it follows that $\underline{U}(\pi) = -\infty$ as well. ■

Corollary 6. *Let $(Bb, \Sigma) \in \hat{\Phi}^M \times \mathbb{R}^{\dim(\hat{\Phi}) \times \dim(\hat{\Phi})}$ be an augmented OLS explanation, and define*

$$\alpha^b(Bb, \Sigma) \in \operatorname*{argmax}_{G : Gb \in \Pi_{\overline{\Phi}}} \operatorname{tr}(G \Sigma B^\top).$$

Then for every targeted policy $\pi : \mathcal{X} \to \Delta(A)$,

$$\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f, \alpha^b(Bb, \Sigma)) \geq \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f, \pi).$$

Proof. Fix any $f^0 = (X^0, Y^0) \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma)$ and let $\mu_{X^0} = (X^0)_{\#}\mu_0$. If π is not μ_{X^0}-almost surely equal to any $\hat{\pi} \in \Pi_{\overline{\Phi}}$, the claim follows from Proposition 8. Otherwise, let $\hat{\pi} = \hat{G}b \in \Pi_{\overline{\Phi}}$ be μ_{X^0}-almost surely equal to π. Then

$$\begin{aligned}
\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f, \alpha^b(Bb, \Sigma)) &= \max_{G : Gb \in \Pi_{\overline{\Phi}}} \operatorname{tr}(G \Sigma B^\top) && \text{(by Proposition 3)} \\
&\geq \operatorname{tr}(\hat{G} \Sigma B^\top) = U(f^0, \hat{\pi}) && \text{(by Proposition 3)} \\
&\geq U(f^0, \pi) \geq \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f, \pi),
\end{aligned}$$

as desired. ■

Lemma 7. *Let $g : \mathcal{X} \to \mathbb{R}$ be continuous. If $g(x^+) > 0$ for some $x^+ \in \mathcal{X}$, then there exists a continuous onto $X^+ : [0, 1] \to \mathcal{X}$ such that $\mathbb{E}_{\omega \sim \mu_0}[g(X^+(\omega))] > 0$.*

Proof. Choose $\varepsilon > 0$ such that $(1 - \varepsilon) g(x^+) + \varepsilon \inf_{x \in \mathcal{X}} g(x) > 0$.

Since every probability measure on $[0, 1]$ has at most countably many atoms, we can choose $\hat{\omega} \in (0, 1)$ such that $\mu_0(\{\hat{\omega}\}) = 0$. By continuity from above, there exists a closed interval $I = [a, b] \subset (0, 1)$ containing $\hat{\omega}$ such that $\mu_0(I) < \varepsilon$.

By Lemma 3, there exists a continuous surjection $q : [0, 1] \to \mathcal{X}$ such that $q(0) = q(1) = x^+$. Define

$$X^+(\omega) := \begin{cases} x^+ & \text{if } \omega \notin I, \\ q\left(\frac{\omega - a}{b - a}\right) & \text{if } \omega \in I. \end{cases}$$

Then X^+ is continuous and onto. Moreover, $\mathbb{E}_{\omega\sim\mu_0}[g(X^+(\omega))] \geq (1-\varepsilon)g(x^+)+\varepsilon \inf_{x\in\mathcal{X}} g(x) > 0$, as desired. ■

Lemma 8 (Mean and Constant OLS)**.** *Fix $\mu \in \Delta([0,1])$. Let $\hat{\Phi}_{\text{const}} := \text{span}\{\mathbf{1}_{\mathcal{X}}\}$, let $\Phi_{\text{const}} := \hat{\Phi}^M_{\text{const}}$, and let $\overline{\Gamma}^{\text{const}}_\mu$ denote the corresponding OLS explainer. Then for every $\phi \in \mathbb{R}^M$ and every $f \in F$,*

$$f \in \mathbb{e}_\mu^{-1}(\phi) \Leftrightarrow f \in (\overline{\Gamma}^{\text{const}}_\mu)^{-1}(\phi_1\mathbf{1}_{\mathcal{X}}, \ldots, \phi_M\mathbf{1}_{\mathcal{X}}).$$

Proof. Fix $f = (X, Y) \in F$. The induced approximation space

$$\Psi^{\text{const}}_X := \{\hat{\phi} \circ X : \hat{\phi} \in \Phi_{\text{const}}\}$$

is exactly the space of constant random vectors in $L^2([0,1], \mu; \mathcal{Y})$. Therefore the orthogonal projection of Y onto Ψ^{const}_X is the constant random vector equal to $\mathbb{e}_\mu(f)$. By the definition of the OLS explainer,

$$\overline{\Gamma}^{\text{const}}_\mu(f) = \big((\mathbb{e}_\mu(f))_1\mathbf{1}_{\mathcal{X}}, \ldots, (\mathbb{e}_\mu(f))_M\mathbf{1}_{\mathcal{X}}\big).$$

Hence $\mathbb{e}_\mu(f) = \phi$ if and only if $\overline{\Gamma}^{\text{const}}_\mu(f) = (\phi_1\mathbf{1}_{\mathcal{X}}, \ldots, \phi_M\mathbf{1}_{\mathcal{X}})$, which is the desired equivalence. ■

Proof of Theorem 3 (i) Identification.

Identification by $\widetilde{\Gamma}^b$. The "if" direction follows immediately from Proposition 3. For the "only if" direction, fix $Bb \in \overline{\Phi}$ and $\pi \notin \Pi_{\overline{\Phi}}$. By Lemma 5, there exists a continuous onto $X^0 : [0,1] \to \mathcal{X}$ such that π is not μ_{X^0}-almost surely equal to any policy in $\hat{\pi} \in \Pi_{\overline{\Phi}}$, where $\mu_{X^0} = (X^0)_{\#}\mu_0$. Let $\Sigma = \mathbb{E}_{\omega\sim\mu_0}[b(X^0(\omega))b(X^0(\omega))^\top]$. Then by Proposition 8, $\inf_{f\in(\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma)} U(f,\pi) = \underline{U}(\pi) = -\infty$. Thus $\widetilde{\Gamma}^b$ does not identify π.

Identification by $\overline{\Gamma}$. Identification follows immediately from Propositions 1 and 2. Suppose toward a contradiction that $\overline{\Gamma}$ exactly identifies $\pi \in \Pi_{\overline{\Phi}}$. Choose a nonconstant $\psi \in \hat{\Phi}$ and let $\phi = (\psi, \ldots, \psi) \in \overline{\Phi}$. Since ψ is not constant, there exist $x_1, x_2 \in \mathcal{X}$ with $\psi(x_1) < \psi(x_2)$; fix $r \in (\psi(x_1), \psi(x_2))$. Applying Lemma 7 to $\psi - r$ (positive at x_2) and to $r - \psi$ (positive at x_1) yields continuous onto processes $X^+, X^- : [0,1] \to \mathcal{X}$ such that

$$\mathbb{E}_{\omega\sim\mu_0}[\psi(X^+(\omega))] > r > \mathbb{E}_{\omega\sim\mu_0}[\psi(X^-(\omega))].$$

Let $f^\pm := (X^\pm, \phi \circ X^\pm) \in F$. Since $Y^\pm = \phi \circ X^\pm \in \Psi_{X^\pm}$, we have $\overline{\Gamma}(f^\pm) = \phi$; moreover $U(f^\pm, \pi) = \mathbb{E}_{\omega\sim\mu_0}[\pi(X^\pm(\omega)) \cdot \phi(X^\pm(\omega))] = \mathbb{E}_{\omega\sim\mu_0}[\psi(X^\pm(\omega))]$. Then $U(f^+, \pi) > r > U(f^-, \pi)$, a contradiction.

Identification by $\mathfrak{e}$. By Lemma 8, $f \in \mathfrak{e}^{-1}(\phi) \Leftrightarrow f \in (\overline{\Gamma}^{\text{const}})^{-1}(\phi_1 \mathbf{1}_{\mathcal{X}}, \ldots, \phi_M \mathbf{1}_{\mathcal{X}})$. Identification then follows from Propositions 1 and 2. For exact identification, suppose that π is constant with $\pi(x) = p$ for each $x \in \mathcal{X}$. Then for any $\phi \in \mathbb{R}^M$ and any $f \in \mathfrak{e}^{-1}(\phi)$, Lemma 6 implies $U(f, \pi) = \mathbb{E}_{\omega \sim \mu_0}[p \cdot Y(\omega)] = p \cdot \phi > -\infty = \underline{U}(\pi)$.

(ii) Informativeness. Since $\overline{\Gamma}$ is obtained from $\widetilde{\Gamma}^b$ by dropping $\Sigma^b(f)$, we have $\overline{\Gamma} = h_1 \circ \widetilde{\Gamma}^b$ for the coordinate projection $h_1 : (Bb, \Sigma) \mapsto Bb$. For $\mathfrak{e}$, observe that since $\mathbf{1} \in \hat{\Phi}$, $\mathbf{1} = c \cdot b$ for some $c \in \mathbb{R}^{\dim \hat{\Phi}}$. Moreover, if $\phi = \overline{\Gamma}(X, Y)$, we have $\mathbb{E}_{\omega \sim \mu_0}[\phi(X(\omega)) - Y(\omega)] = 0$. Then if $\widetilde{\Gamma}^b(f) = (Bb, \Sigma)$, we have

$$\begin{aligned} \mathfrak{e}(f) = \mathbb{E}_{\omega \sim \mu_0}[Bb(X(\omega))] &= \mathbb{E}_{\omega \sim \mu_0}[\sum_a (B_{a\cdot} b(X(\omega)))(c \cdot b(X(\omega)))] \\ &= \sum_a B_{a\cdot} \mathbb{E}_{\omega \sim \mu_0}[b(X(\omega)) b(X(\omega))^\intercal] c = \sum_a B_{a\cdot} \Sigma c = B \Sigma c. \end{aligned} \tag{9}$$

Hence $\mathfrak{e} = h_2 \circ \widetilde{\Gamma}^b$ for the map $h_2 : (Bb, \Sigma) \mapsto B\Sigma c$.

(iii) Robust usefulness. Take $\alpha^* = \alpha^b$, where $\alpha^b : \hat{\Phi}^M \times \mathbb{R}^{\dim \hat{\Phi} \times \dim \hat{\Phi}} \to \Pi$ is defined by

$$\alpha^b(Bb, \Sigma) \in \operatorname*{argmax}_{G : Gb \in \Pi_{\overline{\Phi}}} \operatorname{tr}(G \Sigma B^\top).$$

It follows from Corollary 6 that for $\alpha^* = \alpha^b$, (1) holds for each $\widetilde{\Gamma} \in \{\overline{\Gamma}, \mathfrak{e}\}$. For (2), we consider each $\widetilde{\Gamma}$ separately.

(2) **holds for $\alpha = \alpha^b$ and $\widetilde{\Gamma} = \overline{\Gamma}$.** Choose a nonconstant $\psi \in \hat{\Phi}$. Since ψ is continuous and $\mathcal{X}$ is connected, $\psi(\mathcal{X})$ is a nondegenerate interval. Since $\mathcal{X} \subseteq \mathbb{R}^K$ is second countable, let $\{O_n\}_{n \geq 1}$ be a countable basis for its relative topology. If $\psi^{-1}(r)$ has nonempty interior, then $O_n \subseteq \psi^{-1}(r)$ for some n; because the same basis element cannot be contained in two distinct level sets, this can occur for at most countably many r. We can therefore choose c in the interior of $\psi(\mathcal{X})$ such that $(\psi - c)^{-1}(0)$ has empty interior. Let $h := \psi - c$. Then h takes both positive and negative values, and since $\mathbf{1} \in \hat{\Phi}$, $h \in \hat{\Phi}$ and we can write $h = \eta \cdot b$ for some $\eta \in \mathbb{R}^{\dim \hat{\Phi}} \setminus \{0\}$. Define $B \in \mathbb{R}^{M \times \dim \hat{\Phi}}$ by

$$B = \begin{bmatrix} \eta^\top & -\eta^\top & 0 & \cdots & 0 \end{bmatrix}^\intercal.$$

Thus $Bb(x) = (h(x), -h(x), 0, \ldots, 0)$.

Now fix any $\tilde{\alpha} : \hat{\Phi}^M \to \Pi$, and let $\pi := \tilde{\alpha}(Bb)$. If $\pi \notin \Pi_{\overline{\Phi}}$, then by Lemma 5, there exists a continuous onto $X^0 : [0, 1] \to \mathcal{X}$ such that π is not μ_{X^0}-almost surely equal to any $\hat{\pi} \in \Pi_{\overline{\Phi}}$, where $\mu_{X^0} := (X^0)_{\#}\mu_0$. Let $\Sigma^0 := \mathbb{E}[b(X^0) b(X^0)^\top]$. By Proposition 8, $\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma^0)} U(f, \pi) = -\infty$; by Proposition 3, $\inf_{f \in (\widetilde{\Gamma}^b)^{-1}(Bb, \Sigma^0)} U(f, \alpha^b(Bb, \Sigma^0)) =$

$\max_{G:Gb\in\Pi_{\overline{\Phi}}} \operatorname{tr}(G\Sigma^0 B^\top)$, so (2) follows.

Now suppose $\pi \in \Pi_{\overline{\Phi}}$. Then since π is continuous, we cannot have $\pi(x) \in \arg\max_{p\in\Pi} p \cdot Bb(x)$ for each $x \in \mathcal{X}$: if it did, then $\pi_1(x) = 1$ whenever $h(x) > 0$ and $\pi_1(x) = 0$ whenever $h(x) < 0$. Since h takes both positive and negative values and $\mathcal{X}$ is connected, continuity of π_1 would imply that $\pi_1(x) \in (0,1)$ for some $x \in \mathcal{X}$. For every such x, we must have $h(x) = 0$; moreover, by continuity, $\pi_1(x) \in (0,1)$ on a neighborhood of x, so that neighborhood is contained in $h^{-1}(0)$. This contradicts the fact that $h^{-1}(0)$ has empty interior.

Hence there is $x^* \in \mathcal{X}$ and $a^* \in A$ such that

$$e_{a^*} \cdot Bb(x^*) - \pi(x^*) \cdot Bb(x^*) > 0.$$

Since π and Bb are continuous, $(e_{a^*} - \pi) \cdot Bb$ is continuous. Then by Lemma 7, there exists a continuous onto X^* such that

$$\mathbb{E}_{\omega\sim\mu_0}[(e_{a^*} - \pi(X^*(\omega))) \cdot Bb(X^*(\omega))] = U((X^*, Bb \circ X^*), e_{a^*}) - U((X^*, Bb \circ X^*), \pi) > 0.$$

Then let $\Sigma^* := \mathbb{E}_{\omega\sim\mu_0}[b(X^*(\omega))b(X^*(\omega))^\top]$; since $\mathbf{1} \in \hat{\Phi}$, there exists $E_{a^*} \in \mathbb{R}^{M\times\dim\hat{\Phi}}$ such that $e_{a^*} = E_{a^*}b$. Then by Proposition 3,

$$\begin{aligned}\inf_{f\in(\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma^*)} U(f, \alpha^b(Bb,\Sigma^*)) &= \max_{G:Gb\in\Pi_{\overline{\Phi}}} \operatorname{tr}(G\Sigma^* B^\top) \\ &\geq \operatorname{tr}(E_{a^*}\Sigma^* B^\top) = U((X^*, Bb\circ X^*), e_{a^*}) \\ &> U((X^*, Bb\circ X^*), \pi) \geq \inf_{f\in(\widetilde{\Gamma}^b)^{-1}(Bb,\Sigma^*)} U(f,\pi),\end{aligned}$$

proving (2) for $\alpha = \alpha^b$, $\widetilde{\Gamma} = \overline{\Gamma}$, and $\phi = (Bb, \Sigma^*)$.

(2) **holds for** $\alpha = \alpha^b$ **and** $\widetilde{\Gamma} = \mathfrak{e}$**.** Fix any decision rule $\tilde{\alpha} : \mathbb{R}^M \to \Pi$ and let $\pi = \tilde{\alpha}(0)$. If $\pi \notin \Pi_{\overline{\Phi}}$, let X^0 be given by Lemma 5; otherwise choose any continuous onto $X^0 : [0,1] \to \mathcal{X}$. Choose some nonconstant $\psi \in \hat{\Phi}$; since ψ is nonconstant and $(X^0)_{\#}\mu_0$ has full support, $\psi \circ X^0$ has positive variance. Let

$$h^0(x) := \psi(x) - \mathbb{E}_{\omega\sim\mu_0}[\psi(X^0(\omega))];$$

since $\mathbf{1} \in \hat{\Phi}$, we have $h^0 \in \hat{\Phi}$, and we can write $(h^0, -h^0, 0, \ldots, 0) = B^0 b$ for some $B^0 \in \mathbb{R}^{M\times\dim\hat{\Phi}}$. Now let $\Sigma^0 := \mathbb{E}[b(X^0)b(X^0)^\top]$; by (9), for any $f \in (\widetilde{\Gamma}^b)^{-1}(B^0 b, \Sigma^0)$, we have

$$\mathfrak{e}(f) = B^0\Sigma^0 c = \sum_a \mathbb{E}_{\omega\sim\mu_0}[B^0_{a\cdot} b(X^0(\omega))b(X^0(\omega))^\intercal c] = \mathbb{E}_{\omega\sim\mu_0}[h^0(X^0(\omega))] - \mathbb{E}_{\omega\sim\mu_0}[h^0(X^0(\omega))] = 0;$$

similarly, for any $f' \in (\widetilde{\Gamma}^b)^{-1}(-B^0 b, \Sigma^0)$, we have $\mathbb{e}(f') = 0$.

If $\pi \notin \Pi_{\overline{\Phi}}$, then (2) holds for $\phi = (B^0 b, \Sigma^0)$ by Proposition 8.

If $\pi \in \Pi_{\overline{\Phi}}$, choose $\sigma \in \{-1, 1\}$ such that $\sigma \mathbb{E}_{\omega \sim \mu_0} \left[(\pi_1(X^0(\omega)) - \pi_2(X^0(\omega))) \, h^0(X^0(\omega)) \right] \leq 0$, and define $\pi^\sigma : \mathcal{X} \to \Delta(A)$ by

$$\pi^\sigma(x) := \left(\frac{1}{2} + \frac{\sigma h^0(x)}{2\|h^0\|_\infty}, \frac{1}{2} - \frac{\sigma h^0(x)}{2\|h^0\|_\infty}, 0, \ldots, 0 \right).$$

Since $|h^0(x)| \leq \|h^0\|_\infty$ for every $x \in \mathcal{X}$, $h^0 \in \hat{\Phi}$, and $\mathbf{1} \in \hat{\Phi}$, $\pi^\sigma \in \Pi_{\overline{\Phi}}$. Moreover,

$$\begin{aligned} U((X^0, \sigma B^0 b \circ X^0), \pi) &= \sigma \mathbb{E}_{\omega \sim \mu_0} \left[\left(\pi_1(X^0(\omega)) - \pi_2(X^0(\omega)) \right) h^0(X^0(\omega)) \right] \leq 0, \\ U((X^0, \sigma B^0 b \circ X^0), \pi^\sigma) &= \mathbb{E}_{\omega \sim \mu_0} \left[\frac{(h^0(X^0(\omega)))^2}{\|h^0\|_\infty} \right] > 0. \end{aligned}$$

Then by Corollary 6,

$$\begin{aligned} \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\sigma B^0 b, \Sigma^0)} U(f, \alpha^b(\sigma B^0 b, \Sigma^0)) \geq U((X^0, \sigma B^0 b \circ X^0), \pi^\sigma) > 0 \geq U((X^0, \sigma B^0 b \circ X^0), \pi) \\ \geq \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\sigma B^0 b, \Sigma^0)} U(f, \pi) \\ = \inf_{f \in (\widetilde{\Gamma}^b)^{-1}(\sigma B^0 b, \Sigma^0)} U(f, \tilde{\alpha}(\mathbb{e}(f))), \end{aligned}$$

where the last equality follows since $\mathbb{e}(f) = 0$ for every $f \in (\widetilde{\Gamma}^b)^{-1}(\sigma B^0 b, \Sigma^0)$. Thus (2) holds for $\widetilde{\Gamma} = \mathbb{e}$, $\alpha = \alpha^b$, and $\phi = (\sigma B^0 b, \Sigma^0)$.

Thus, with $\alpha^* = \alpha^b$, both (1) and (2) hold for each $\widetilde{\Gamma} \in \{\overline{\Gamma}, \mathbb{e}\}$, proving (iii). ■

Proof of Proposition 4 Define $f^\dagger = (X^\dagger, Y^\dagger)$ by $X^\dagger(\omega) := \omega$ and $Y^\dagger(\omega) := (\omega, 1-\omega)$. Then $f^\dagger \in F$. Fix any finite-dimensional subspace $\hat{\Phi} \subseteq C([0,1])$ containing the constant functions, let $\overline{\Gamma}$ be the corresponding OLS explainer with $\overline{\Phi} = \hat{\Phi}^2$, and write $\phi = \overline{\Gamma}(f^\dagger)$. Then by symmetry, $(\phi_2, \phi_1) = \overline{\Gamma}(X^\dagger, 1 - Y^\dagger)$. Then since $\overline{\Gamma}$ is linear in Y, we have $(\phi_1 + \phi_2, \phi_1 + \phi_2) = \overline{\Gamma}(X^\dagger, (\mathbf{1}, \mathbf{1}))$. But $\mathbf{1} \in \hat{\Phi}$, so we must have $\phi_1 + \phi_2 = \mathbf{1}$.

Since $\phi \circ X^\dagger = \phi$ is the orthogonal projection of $Y^\dagger$ onto $\Psi_{X^\dagger}$, and $(\mathbf{1}, \mathbf{0}) \in \overline{\Phi}$, we have

$$\int_0^1 \phi_1(\omega) \, \mathrm{d}\omega = \int_0^1 Y_1^\dagger(\omega) \, \mathrm{d}\omega = \frac{1}{2}.$$

Since ϕ_1 is continuous, there exists $x^* \in [0,1]$ such that $\phi_1(x^*) = 1/2$. Then since $\phi_1 = 1 - \phi_2$, we have $\phi(x^*) = (1/2, 1/2)$.

For any $\pi \in \Pi_{\overline{\Phi}}$, Proposition 1 implies

$$\inf_{\hat{f} \in \overline{\Gamma}^{-1}(\overline{\Gamma}(f^\dagger))} U(\hat{f}, \pi) = \inf_{x \in [0,1]} \pi(x) \cdot \phi(x) \leq \pi(x^*) \cdot \phi(x^*) = \frac{1}{2},$$

while if $\pi \notin \Pi_{\overline{\Phi}}$, we have $\inf_{\hat{f} \in \overline{\Gamma}^{-1}(\overline{\Gamma}(f^\dagger))} U(\hat{f}, \pi) = \underline{U}(\pi) = -\infty$ by Proposition 2 and Lemma 6. Hence

$$\sup_{\pi \in \Pi} \inf_{\hat{f} \in \overline{\Gamma}^{-1}(\overline{\Gamma}(f^\dagger))} U(\hat{f}, \pi) \leq \frac{1}{2}.$$

By contrast, we have

$$\overline{U}(f^\dagger) = \int_0^1 \max\{x, 1-x\} \, \mathrm{d}x = \frac{3}{4},$$

as desired. ■

Given a continuous $c : \mathcal{X} \to \mathcal{Y}$, we construct $B_m c$, the tensor-product Bernstein approximation to c of order m, as follows. Let $B = \prod_{i=1}^K [a_i, b_i] \supseteq \mathcal{X}$ be a box containing $\mathcal{X}$. Since $\Delta(A)$ is convex, Dugundji's extension theorem (Dugundji 1951) implies that every continuous map $c : \mathcal{X} \to \Delta(A)$ admits a continuous extension $\widetilde{c} : B \to \Delta(A)$. For each $x \in \mathcal{X}$, write $t_i(x) := (x_i - a_i)/(b_i - a_i)$. Then define $B_m c : B \to \Delta(A)$ by

$$B_m c(x) := \sum_{j \in \{0,\ldots,m\}^K} \widetilde{c}\left(a_1 + \frac{j_1}{m}(b_1 - a_1), \ldots, a_K + \frac{j_K}{m}(b_K - a_K)\right) \prod_{k=1}^K \binom{m}{j_k} t_k(x)^{j_k} (1 - t_k(x))^{m - j_k}.$$

Lemma 9. *Let $\mathcal{X} \subseteq \mathbb{R}^K$ be compact and let $c : \mathcal{X} \to \Delta(A)$ be continuous. Then for every $m \geq 1$, $B_m c(x) \in \Delta(A)$ for every $x \in \mathcal{X}$, and*

$$\sup_{x \in \mathcal{X}} \|B_m c(x) - c(x)\|_\infty \to 0$$

as $m \to \infty$.

Proof. For each $x \in \mathcal{X}$, the Bernstein weights

$$\left\{ \prod_{k=1}^K \binom{m}{j_k} t_k(x)^{j_k} (1 - t_k(x))^{m - j_k} \right\}_{j \in \{0,\ldots,m\}^K}$$

are nonnegative and sum to one. Since $\widetilde{c}(y) \in \Delta(A)$ for all $y \in B$, it follows that $B_m c(x)$ is a convex combination of elements of $\Delta(A)$; since $\Delta(A)$ is convex, it follows that $B_m c(x) \in \Delta(A)$.

Because $\widetilde{c}$ is continuous on the box B, the multivariate Bernstein approximation theorem

(after the affine change of variables from B to $[0,1]^K$; see, e.g., Lorentz 1986) implies that

$$\|B_m c - c\|_\infty \leq \sup_{x \in B} \|B_m \widetilde{c}(x) - \widetilde{c}(x)\|_\infty \to 0.$$

as desired. ■

Proof of Proposition 5 For each explanation $\phi \in \Gamma_n(F)$, fix an arbitrary $f_\phi \in \Gamma_n^{-1}(\phi)$ and define the decision rule

$$\alpha_n(\phi) \in \operatorname*{argmax}_{\pi \in \Pi_n^*} U(f_\phi, \pi);$$

the choice of $\alpha_n(\phi)$ is well-defined because, by hypothesis, $U(\cdot, \pi)$ is constant on $\Gamma_n^{-1}(\phi)$ for every $\pi \in \Pi_n^*$. Then for every $f \in F$,

$$\inf_{\hat{f} \in \Gamma_n^{-1}(\Gamma_n(f))} U(\hat{f}, \alpha_n(\Gamma_n(f))) = U(f, \alpha_n(\Gamma_n(f))) = \sup_{\pi \in \Pi_n^*} U(f, \pi) \to \overline{U}(f),$$

where the first equality uses exact identification of $\alpha_n(\Gamma_n(f)) \in \Pi_n^*$, the second uses the definition of α_n at $\Gamma_n(f)$, and the limit is the hypothesis on $\{\Pi_n^*\}$.

Proof of Theorem 4 (Augmented OLS Is Asymptotically Robustly Perfect) Fix $f \in F$. For each n, Proposition 3 implies that for each $\pi \in \Pi_{\overline{\Phi}_n}$ and $\hat{f} \in (\widetilde{\Gamma}_n^b)^{-1}(\widetilde{\Gamma}_n^b(f))$, we have $U(\hat{f}, \pi) = U(f, \pi)$. Therefore

$$\inf_{\hat{f} \in (\widetilde{\Gamma}_n^b)^{-1}(\widetilde{\Gamma}_n^b(f))} U(\hat{f}, \alpha_n^b(\widetilde{\Gamma}_n^b(f))) = \sup_{\pi \in \Pi_{\overline{\Phi}_n}} U(f, \pi).$$

Let $\mu_f := (X)_\# \mu_0$ be the distribution of covariates induced by f, and let $B_f := \sup_{\omega \in [0,1]} \|Y(\omega)\|_\infty < \infty$. Fix $\varepsilon > 0$. By the definition of $\overline{U}(f)$, there exists a policy $\pi_{f,\varepsilon}^* : \mathcal{X} \to \Delta(A)$ such that $\overline{U}(f) - U(f, \pi_{f,\varepsilon}^*) < \frac{\varepsilon}{2}$. By Lusin's theorem, there exists a continuous $c_\varepsilon : \mathcal{X} \to \Delta(A)$ such that

$$\|c_\varepsilon - \pi_{f,\varepsilon}^*\|_{L^1(\mu_f)} < \frac{\varepsilon}{4B_f + 1}.$$

For each sufficiently large n, let $m_n := \lfloor n/K \rfloor$ and $\pi_{n,\varepsilon} := B_{m_n} c_\varepsilon$. By Lemma 9, $\|\pi_{n,\varepsilon} - c_\varepsilon\|_\infty \to 0$. Moreover, each coordinate of $\pi_{n,\varepsilon}$ is the restriction to $\mathcal{X}$ of a polynomial of total degree at most $Km_n \leq n$, so $\pi_{n,\varepsilon} \in \Pi_{\overline{\Phi}_n}$. Hence, for all sufficiently large n,

$$\|\pi_{n,\varepsilon} - c_\varepsilon\|_{L^1(\mu_f)} \leq \|\pi_{n,\varepsilon} - c_\varepsilon\|_\infty < \frac{\varepsilon}{4B_f + 1}.$$

It follows that $\|\pi_{n,\varepsilon} - \pi_{f,\varepsilon}^*\|_{L^1(\mu_f)} < \frac{2\varepsilon}{4B_f+1}$ for all sufficiently large n.

For those n,

$$\begin{aligned}\overline{U}(f) - \sup_{\pi \in \Pi_{\overline{\Phi}_n}} U(f,\pi) &\leq \overline{U}(f) - U(f,\pi^*_{f,\varepsilon}) + U(f,\pi^*_{f,\varepsilon}) - U(f,\pi_{n,\varepsilon}) \\ &< \frac{\varepsilon}{2} + U(f,\pi^*_{f,\varepsilon}) - U(f,\pi_{n,\varepsilon}) \\ &= \frac{\varepsilon}{2} + \mathbb{E}_{\omega\sim\mu_0}\left[Y(\omega)\cdot\left(\pi^*_{f,\varepsilon}(X(\omega)) - \pi_{n,\varepsilon}(X(\omega))\right)\right] \\ &\leq \frac{\varepsilon}{2} + B_f\|\pi^*_{f,\varepsilon} - \pi_{n,\varepsilon}\|_{L^1(\mu_f)} \\ &< \frac{\varepsilon}{2} + \frac{2B_f\varepsilon}{4B_f+1} < \varepsilon.\end{aligned}$$

Since $\sup_{\pi\in\Pi_{\overline{\Phi}_n}} U(f,\pi) \leq \overline{U}(f)$ always, this proves

$$\sup_{\pi\in\Pi_{\overline{\Phi}_n}} U(f,\pi) \to \overline{U}(f).$$

Combining this with the first display in the proof yields

$$\inf_{\hat{f}\in(\widetilde{\Gamma}^b_n)^{-1}(\widetilde{\Gamma}^b_n(f))} U(\hat{f}, \alpha^b_n(\widetilde{\Gamma}^b_n(f))) \to \overline{U}(f),$$

as desired. ■

Proposition 9. *For any weak-∗ open set $\mathcal{M} \subseteq \Delta([0,1])$ containing μ_0, there exists $\mu \in \mathcal{M}$ such that for all $\phi^* \in \overline{\Phi}$, $\pi : \mathcal{X} \to \Delta(A)$, and $z \in \mathbb{R}$, there exists $f = (X,Y) \in \overline{\Gamma}^{-1}_\mu(\phi^*)$ such that*

$$\mathbb{E}_{\omega\sim\mu_0}[\pi(X(\omega))\cdot Y(\omega)] = z.$$

Proof. Let $\{\psi_j\}_{j=1}^{\dim(\hat{\Phi})}$ be a basis for $\hat{\Phi}$. By Lemma 2, we can choose $\{x_i\}_{i=1}^{\dim(\hat{\Phi})} \subset \mathcal{X}$ such that

$$\Psi = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_{\dim(\hat{\Phi})}(x_1) \\ \vdots & \ddots & \vdots \\ \psi_1(x_{\dim(\hat{\Phi})}) & \cdots & \psi_{\dim(\hat{\Phi})}(x_{\dim(\hat{\Phi})}) \end{bmatrix}$$

has full rank. Since every probability measure on $[0,1]$ has at most countably many atoms, we can choose $\hat{\omega} \in (0,1)$ such that $\mu_0(\{\hat{\omega}\}) = 0$. By continuity from above, there exists a closed interval $[a,b] \subset (0,1)$ containing $\hat{\omega}$ such that $\mu_0([a,b]) < \frac{1}{2}$; let $J = [0,1] \setminus [a,b]$. Then $\mu_0(J) \in (\frac{1}{2},1)$. Choose distinct points $\{\omega_i\}_{i=1}^{\dim(\hat{\Phi})}$ with $a = \omega_0 < \omega_1 < \cdots < \omega_{\dim(\hat{\Phi})} < \omega_{\dim(\hat{\Phi})+1} = b$ such that $\mu_0(\{\omega_i\}) = 0$ for every $i = 1,\ldots,\dim(\hat{\Phi})$.

Since $\mathcal{M}$ is open in the weak-$*$ topology, there exists $\lambda \in (0,1)$ such that

$$\mu := \lambda\mu_0 + (1-\lambda)\sum_{i=1}^{\dim(\hat{\Phi})} \frac{1}{\dim(\hat{\Phi})}\delta_{\omega_i} \in \mathcal{M}.$$

Now choose $x_0 \in \mathcal{X}$. For each $i = 1, \ldots, \dim(\hat{\Phi}) + 1$, Lemma 3 gives a continuous surjection $q_i : [0,1] \to \mathcal{X}$ such that $q_i(0) = x_{i-1}$ and $q_i(1) = x_i$ for $i = 1, \ldots, \dim(\hat{\Phi})$, and $q_i(1) = x_0$ for $i = \dim(\hat{\Phi}) + 1$. Define a continuous onto input process $\hat{X} : [0,1] \to \mathcal{X}$ by

$$\hat{X}(\omega) := \begin{cases} x_0 & \text{if } \omega \in J, \\ q_i\left(\frac{\omega-\omega_{i-1}}{\omega_i-\omega_{i-1}}\right) & \text{if } \omega \in [\omega_{i-1}, \omega_i] \text{ for some } i = 1, \ldots, \dim(\hat{\Phi}) + 1. \end{cases}$$

Let $v = \begin{bmatrix} \psi_1(x_0) & \cdots & \psi_{\dim(\hat{\Phi})}(x_0) \end{bmatrix}^\intercal$, and define

$$s := -\frac{\lambda\mu_0(J)\dim(\hat{\Phi})}{1-\lambda}\Psi^{-\intercal}v.$$

Let $r : [0,1] \to \mathbb{R}$ be the bounded Borel measurable function

$$r(\omega) := \begin{cases} 1 & \text{if } \omega \in J, \\ s_i & \text{if } \omega = \omega_i \text{ for some } i \in \{1, \ldots, \dim(\hat{\Phi})\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $r(\omega) = 1$ for every $\omega \in J$, $r(\omega) = 0$ for μ_0-almost every $\omega \notin J$, and $r(\omega_i) = s_i$ for each $0 < i \leq \dim(\hat{\Phi})$.

Now fix $\phi^* \in \overline{\Phi}$, $\pi : \mathcal{X} \to \Delta(A)$, and $z \in \mathbb{R}$. Define

$$\alpha := \frac{z - \mathbb{E}_{\omega\sim\mu_0}[\pi(\hat{X}(\omega)) \cdot \phi^*(\hat{X}(\omega))]}{\mu_0(J)\|\pi(x_0)\|^2},$$

and let $\hat{f} := (\hat{X}, \hat{Y})$, where $\hat{Y} := \phi^* \circ \hat{X} + \alpha r\pi(x_0)$. Because $\mu_0(\{\omega_i\}) = 0$ for all i, we have $r(\omega) = 0$ for μ_0-almost every $\omega \notin J$. Therefore

$$\mathbb{E}_{\omega\sim\mu_0}[\pi(\hat{X}(\omega)) \cdot \hat{Y}(\omega)] = \mathbb{E}_{\omega\sim\mu_0}[\pi(\hat{X}(\omega)) \cdot \phi^*(\hat{X}(\omega))] + \alpha\mu_0(J)\pi(x_0) \cdot \pi(x_0) = z.$$

It remains to show that $\overline{\Gamma}_\mu(f) = \phi^*$. For each basis function ψ_j, we have

$$\int_{[0,1]} r(\omega)\psi_j(\hat{X}(\omega))\,\mathrm{d}\mu(\omega) = \lambda\mu_0(J)\psi_j(x_0) + \frac{1-\lambda}{\dim(\hat{\Phi})}\sum_{i=1}^{\dim(\hat{\Phi})} s_i\psi_j(x_i) = \lambda\mu_0(J)v_j + \frac{1-\lambda}{\dim(\hat{\Phi})}(\Psi^\intercal s)_j = 0.$$

Since $\{\psi_j\}_{j=1}^{\dim(\hat{\Phi})}$ is a basis for $\hat{\Phi}$, it follows that $\int_{[0,1]} r(\omega)\psi(\bar{x}(\omega))\,\mathrm{d}\mu(\omega) = 0$ for every $\psi \in \hat{\Phi}$. Consequently, for every $\phi \in \overline{\Phi}$,

$$\left\langle rw, \phi \circ \hat{X} \right\rangle_\mu = \sum_{m=1}^{M} w_m \int_{[0,1]} r(\omega)\phi_m(\hat{X}(\omega))\,\mathrm{d}\mu(\omega) = 0.$$

Thus rw is orthogonal to $\Psi_{\hat{X}}$. Since $\phi^* \circ \hat{X} \in \Psi_{\hat{X}}$, the orthogonal projection of $\hat{Y} = \phi^* \circ \hat{X} + \alpha rw$ onto $\Psi_{\hat{X}}$ is $\phi^* \circ \hat{X}$. By the definition of the OLS explainer, this means $\overline{\Gamma}_\mu(\hat{f}) = \phi^*$. ■

Proposition 10. *In an untargeted utilitarian decision problem, for any open $\mathcal{M} \subseteq \Delta([0,1])$ containing μ_0, there exists $\mu \in \mathcal{M}$ such that explanation with $\overline{\Gamma}_\mu$ is useless for robust decisions. That is, for every $\phi \in \Phi$ and every $\pi \in \Delta(A)$,*

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

Proof. Let $\mu \in \mathcal{M}$ be given by Proposition 9. Fix $\phi^* \in \Phi$ and $\pi \in \Delta(A)$. Since $\overline{\Gamma}_\mu^{-1}(\phi^*) \subseteq F$, we have

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

For any $z \in \mathbb{R}$, Proposition 9 (with any $x_0 \in \mathcal{X}$ and constant π) yields $f = (X,Y) \in \overline{\Gamma}_\mu^{-1}(\phi^*)$ such that

$$U(f,\pi) = \pi \cdot \mathbb{E}_{\omega\sim\mu_0}[Y(\omega)] = z.$$

Hence

$$\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f,\pi) \leq z \quad \text{for every } z \in \mathbb{R},$$

so $\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f,\pi) = -\infty$. The previous inequality therefore implies $\underline{U}(\pi) = -\infty$ as well, and thus

$$\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f,\pi) = \underline{U}(\pi),$$

as claimed. ■

Proof of Theorem 5 (Sample Means Are Not Robustly Useful) Follows immediately

from Lemma 8 and Proposition 10.

■

Proof of Theorem 6 (Sampled OLS Is Useless in Targeted Problems) Fix an open set $\mathcal{M} \subseteq \Delta([0,1])$ containing μ_0, and let $\mu \in \mathcal{M}$ be given by Proposition 9. Fix $\phi^* \in \overline{\Phi}$ and a targeted policy $\pi : \mathcal{X} \to \Delta(A)$.

Since $\overline{\Gamma}_\mu^{-1}(\phi^*) \subseteq F$, we have

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

For any $z \in \mathbb{R}$, Proposition 9 (with $w = \pi(x_0)$) yields $f = (X, Y) \in \overline{\Gamma}_\mu^{-1}(\phi^*)$ such that

$$U(f, \pi) = \mathbb{E}_{\omega \sim \mu_0} \left[\pi(X(\omega)) \cdot Y(\omega) \right] = z.$$

Hence

$$\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f, \pi) \leq z \quad \text{for every } z \in \mathbb{R},$$

so $\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f, \pi) = -\infty$. The previous inequality therefore implies $\underline{U}(\pi) = -\infty$ as well, and thus

$$\inf_{f \in \overline{\Gamma}_\mu^{-1}(\phi^*)} U(f, \pi) = \underline{U}(\pi),$$

which is exactly the claim. ■

Proof of Proposition 6 (Reweighting Restores Population Objects) Fix $f = (X, Y) \in F$. Since μ satisfies sample selection on observables, we have

$$\mathfrak{e}_\mu^\rho(f) = \int_{[0,1]} \rho(X(\omega)) Y(\omega) \, \mathrm{d}\mu(\omega) = \int_{[0,1]} \frac{d\mu_0}{d\mu}(\omega) Y(\omega) \, \mathrm{d}\mu(\omega) = \int_{[0,1]} Y(\omega) \, \mathrm{d}\mu_0(\omega) = \mathfrak{e}(f).$$

Since μ satisfies sample selection on observables, we have

$$\begin{aligned} \mathbb{E}_{\omega \sim \mu} \left[\rho(X(\omega)) \|Y(\omega) - \phi(X(\omega))\|^2 \right] &= \int \frac{d\mu_0}{d\mu}(\omega) \|Y(\omega) - \phi(X(\omega))\|^2 \, \mathrm{d}\mu(\omega) \\ &= \mathbb{E}_{\omega \sim \mu_0} \left[\|Y(\omega) - \phi(X(\omega))\|^2 \right]. \end{aligned}$$

Therefore the weighted least-squares objective defining $\overline{\Gamma}_\mu^\rho(f)$ coincides exactly with the population least-squares objective defining $\overline{\Gamma}(f)$, so $\overline{\Gamma}_\mu^\rho(f) = \overline{\Gamma}(f)$.

Finally, since μ satisfies sample selection on observables, we have

$$\Sigma^{b,\rho}(f) = \mathbb{E}_{\omega\sim\mu}\left[\rho(X(\omega))b(X(\omega))b(X(\omega))^\top\right] = \int \frac{d\mu_0}{d\mu}(\omega)b(X(\omega))b(X(\omega))^\top \, \mathrm{d}\mu(\omega)$$
$$= \mathbb{E}_{\omega\sim\mu_0}\left[b(X(\omega))b(X(\omega))^\top\right] = \Sigma^b(f).$$

This proves the claim.

Proof of Corollary 2 (Reweighted Means Are Robustly Perfect) Proposition 6 implies that $\mathbb{e}^\rho_\mu(f) = \mathbb{e}(f)$ for every $f \in F$. Hence Theorem 1 applies verbatim.

Proof of Corollary 3 (Weighted Least Squares Is Robustly Useful) Proposition 6 implies that $\overline{\Gamma}^\rho_\mu(f) = \overline{\Gamma}(f)$ for every $f \in F$. The claim therefore follows immediately from Theorem 2.

Proof of Corollary 4 (Augmented WLS Is Asymptotically Robustly Perfect) For each $n \geq 1$, Proposition 6 implies that

$$\widetilde{\Gamma}^{b,\rho}_{\mu,n}(f) = \widetilde{\Gamma}^b_n(f)$$

for every $f \in F$. Hence, Theorem 4 applies verbatim.

Proof of Proposition 7 Choose $\hat{f} = (\hat{X}, \hat{Y}) \in \Gamma^{-1}(\phi^*)$, and let $H := B_b([0,1]; \mathcal{Y})$. Since F contains all bounded Borel measurable DGPs, $(\hat{X}, Y) \in F$ for every $Y \in H$. Define

$$T_{\hat{X}} : H \to \Phi, \qquad T_{\hat{X}}(Y) := \Gamma(\hat{X}, Y) - \Gamma(\hat{X}, 0).$$

By affineness in outcomes, $T_{\hat{X}}$ is linear. Hence, by the first isomorphism theorem, $\dim(H/\ker T_{\hat{X}}) = \dim(\mathrm{Im}(T_{\hat{X}})) \leq \dim(\Gamma(F))$. Now let

$$\mathcal{M}_{z,q,\phi^*}(\hat{X}) := \left\{\mu \in \Delta([0,1]) \,\middle|\, \mathbb{E}_{\omega\sim\mu}[q(\omega)\cdot Y(\omega)] \neq z \text{ for all } Y \in H \text{ with } (\hat{X}, Y) \in \Gamma^{-1}(\phi^*)\right\}$$

It is immediate that $\mathcal{M}_{z,q,\phi^*} \subseteq \mathcal{M}_{z,q,\phi^*}(\hat{X})$. Indeed, if $\mu \notin \mathcal{M}_{z,q,\phi^*}(\hat{X})$, then there exists $Y \in H$ with $\Gamma(\hat{X}, Y) = \phi^*$ and $\mathbb{E}_{\omega\sim\mu}[q(\hat{X}(\omega))Y(\omega)] = z$. But then $(\hat{X}, Y) \in \Gamma^{-1}(\phi^*)$, so $\mu \notin \mathcal{M}_{z,q,\phi^*}$. It therefore suffices to show that $\dim(\mathcal{M}_{z,q,\phi^*}(\hat{X})) < \dim(\Gamma(F))$.

Suppose toward a contradiction that $\dim(\mathcal{M}_{z,q,\phi^*}(\hat{X})) \geq \dim(\Gamma(F))$. Then there exists a (without loss, finite) affinely independent set $S \subseteq \mathcal{M}_{z,q,\phi^*}(\hat{X})$ with $|S| + 1 \geq \dim(\Gamma(F))$. Since $S \subset \Delta([0,1])$, S is linearly dependent: if $\sum_{\mu\in S}\alpha_\mu\mu = 0$, then $\sum_{\mu\in S}\alpha_\mu\mu([0,1]) = \sum_{\mu\in S}\alpha_\mu = 0$. For each $\mu \in S$, define the linear functional

$$e_{\mu,q} : H \to \mathbb{R}, \qquad e_{\mu,q}(Y) := \mathbb{E}_{\omega\sim\mu}[q(\hat{X}(\omega))\cdot Y(\omega)].$$

We claim that the family $\{e_{\mu,q}\}_{\mu\in S}$ is linearly independent. Indeed, if $\sum_{\mu\in S} c_\mu e_{\mu,q} = 0$, then for every bounded Borel measurable scalar function $s : [0,1] \to \mathbb{R}$, the function

$$Y_s(\omega) := s(\omega)\frac{q(\hat{X}(\omega))}{\|q(\hat{X}(\omega))\|^2}$$

is an element of H, and

$$0 = \sum_{\mu\in S} c_\mu e_{\mu,q}(Y_s) = \sum_{\mu\in S} c_\mu \mathbb{E}_{\omega\sim\mu}[s(\omega)].$$

Since this holds for every bounded Borel measurable s, taking $s = \mathbf{1}_B$ for each Borel set $B \subseteq [0,1]$ shows that $\sum_{\mu\in S} c_\mu \mu = 0$ as a signed Borel measure. Because S is linearly independent, each c_μ must be zero.

Next we claim that for every $\mu \in S$, $\ker(T_{\hat{X}}) \subseteq \ker(e_{\mu,q})$. Fix $\mu \in S$, and suppose that some $g \in \ker(T_{\hat{X}})$ satisfies $e_{\mu,q}(g) \neq 0$. Let $\alpha = \frac{z - e_{\mu,q}(\hat{Y})}{e_{\mu,q}(g)}$; then $e_{\mu,q}(\hat{Y} + \alpha g) = z$. But since $T_{\hat{X}}(g) = 0$, affineness in outcomes implies $T_{\hat{X}}(\hat{Y} + \alpha g) = T_{\hat{X}}(\hat{Y})$ and hence $\Gamma(\hat{X}, \hat{Y} + \alpha g) = \Gamma(\hat{X}, \hat{Y}) = \phi^*$. Then $\mu \notin \mathcal{M}_{z,q,\phi^*}(\hat{X})$, a contradiction.

It follows that each $e_{\mu,q}$ factors through the quotient $H/\ker(T_{\hat{X}})$: that is, for each $\mu \in S$ there exists a linear functional $\widetilde{e}_{\mu,q} : H/\ker(T_{\hat{X}}) \to \mathbb{R}$ such that $e_{\mu,q} = \widetilde{e}_{\mu,q} \circ Q$, where $Q : H \to H/\ker(T_{\hat{X}})$ is the quotient map. Since the functionals $\{e_{\mu,q}\}_{\mu\in S}$ are linearly independent, so are $\{\widetilde{e}_{\mu,q}\}_{\mu\in S}$. But $\dim(H/\ker(T_{\hat{X}})) \leq \dim(\Gamma(F))$, so its dual space $(H/\ker(T_{\hat{X}}))^*$ also has dimension at most $\dim(\Gamma(F))$. Then $|S| \leq \dim(\Gamma(F))$, a contradiction.

Therefore $\dim(\mathcal{M}_{z,q,\phi^*}) \leq \dim(\mathcal{M}_{z,q,\phi^*}(\hat{X})) < \dim(\Gamma(F))$, as desired. ■

Proof of Theorem 7 (Affine Explainers are Not Robustly Useful with Ambiguity Aversion) Fix $\phi \in \Gamma(F)$, and for each $\pi \in \Pi$ and $z \in \mathbb{R}$, let $\mathcal{M}_{z,\pi,\phi}$ be as in (7) (identifying each $\pi \in \Delta(A)$ with the constant map in the case of an untargeted problem). By hypothesis, $\dim(\Gamma(F)) \leq \dim(\mathcal{M})$. Then by Proposition 7, for each $\pi \in \Pi$ and $z \in \mathbb{R}$, $\dim(\mathcal{M}_{z,\pi,\phi}) < \dim(\Gamma(F)) \leq \dim(\mathcal{M})$. Therefore $\mathcal{M} \nsubseteq \mathcal{M}_{z,\pi,\phi}$, so there exist $\mu_z \in \mathcal{M}$ and $f^z = (X^z, Y^z) \in \Gamma^{-1}(\phi)$ such that $\mathbb{E}_{\omega\sim\mu_z}[\pi(X^z(\omega)) \cdot Y^z(\omega)] = z$. Hence

$$\inf_{f\in\Gamma^{-1}(\phi)} U_{\mathcal{M}}(f,\pi) = \inf_{\substack{f\in\Gamma^{-1}(\phi)\\ \mu\in\mathcal{M}}} \mathbb{E}_{\omega\sim\mu}[\pi(X(\omega)) \cdot Y(\omega)] \leq z.$$

Since $z \in \mathbb{R}$ was arbitrary, it follows that $\inf_{f\in\Gamma^{-1}(\phi)} U_{\mathcal{M}}(f,\pi) = -\infty \leq \inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x) \cdot y$.

On the other hand, because $\Gamma^{-1}(\phi) \subseteq F$, we have $\underline{U}_{\mathcal{M}}(\pi) \leq \inf_{f\in\Gamma^{-1}(\phi)} U_{\mathcal{M}}(f,\pi)$. Also,

for every $f = (X, Y) \in F$ and every $\mu \in \mathcal{M}$,

$$\mathbb{E}_{\omega\sim\mu}[\pi(X(\omega)) \cdot Y(\omega)] \geq \inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x) \cdot y,$$

and therefore

$$\inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x) \cdot y \leq \underline{U}_{\mathcal{M}}(\pi).$$

Then we have

$$\inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x) \cdot y \leq \underline{U}_{\mathcal{M}}(\pi) \leq \inf_{f\in\Gamma^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) \leq \inf_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \pi(x) \cdot y.$$

and so the quantities must be equal. Taking maxima over Π yields (6). ■

Proof of Corollary 5 (Affine Explainers are Not Robustly Useful in Rawlsian Problems) Recall that $R(f, \pi) := U_{\underline{\mathcal{M}}}(f, \pi)$ for every f and π, and hence $\underline{R}(\pi) = \underline{U}_{\underline{\mathcal{M}}}(\pi)$. The Dirac measures $\underline{\mathcal{M}} := \{\delta_\omega\}_{\omega\in[0,1]}$ are affinely independent, so $\underline{\mathcal{M}}$ is infinite-dimensional; in particular $\dim(\underline{\mathcal{M}}) > \dim(\Gamma(F))$ since $\Phi \supseteq \Gamma(F)$ is finite-dimensional. Applying Theorem 7 with $\mathcal{M} = \underline{\mathcal{M}}$ therefore yields

$$\inf_{f\in\Gamma^{-1}(\phi)} R(f, \pi) = \underline{R}(\pi) \qquad \text{for all } \phi \in \Gamma(F) \text{ and } \pi \in \Pi,$$

and (8) follows by taking maxima over Π. ■

Proof of Theorem 8 (Action Certificates Are Robustly Useful) Fix an explanation $\phi \in \mathbb{R}^A$ and a policy $\pi \in \Delta(A)$. Then for any $f \in \underline{\Gamma}^{-1}_{A,\mathcal{M}}(\phi)$, we have

$$U_{\mathcal{M}}(f, \pi) = \inf_{\mu\in\mathcal{M}} \sum_{a\in A} \pi_a \, \mathbb{E}_{\omega\sim\mu}[Y_a(\omega)] \geq \sum_{a\in A} \pi_a \phi_a = \pi \cdot \phi,$$

and therefore $\inf_{f\in(\underline{\Gamma}_{A,\mathcal{M}})^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) \geq \pi \cdot \phi$. For the reverse inequality, choose any continuous onto $X : [0, 1] \to \mathcal{X}$ and let $Y(\omega) = \phi$ for each $\omega \in [0, 1]$. Then $(X, Y) \in \underline{\Gamma}^{-1}_{A,\mathcal{M}}(\phi)$, and $U_{\mathcal{M}}((X, Y), \pi) = \pi \cdot \phi$; consequently, $\inf_{f\in(\underline{\Gamma}_{A,\mathcal{M}})^{-1}(\phi)} U_{\mathcal{M}}(f, \pi) \leq \pi \cdot \phi$.

Then for every policy π,

$$\inf_{f\in(\underline{\Gamma}_{A,\mathcal{M}})^{-1}(\phi)} U_{\mathcal{M}}(f, \alpha^*(\phi)) = \max_{\pi'\in\Delta(A)} \pi' \cdot \phi \geq \pi \cdot \phi = \inf_{f\in(\underline{\Gamma}_{A,\mathcal{M}})^{-1}(\phi)} U_{\mathcal{M}}(f, \pi),$$

with strict inequality for $\phi = e_{a^*}$ for any $a^* \in \arg\min_a \pi_a$. It follows that $\underline{\Gamma}_{A,\mathcal{M}}$ is robustly useful. ■

Proof of Theorem 9 (Partition Certificates Are Robustly Useful) Fix an explanation $\phi \in \underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(F)$. Then ϕ is constant on each $C \in \mathcal{C}$; for convenience write its value on C as $\phi(C)$. Say that a cell C is active for ϕ if $\phi(C)_a < \infty$ for some $a \in A$; let $\mathcal{A}(\phi)$ denote the set of active cells. By definition of F, for each $f = (X, Y) \in F$, Y is bounded; then C is active for ϕ if and only if for every $(X, Y) \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$, $\mu(X^{-1}(C)) > 0$ for some $\mu \in \mathcal{M}$.

Define

$$W(\phi) := \inf_{(X,Y)\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} \inf_{\mu\in\mathcal{M}} \sum_{C\in\mathcal{A}(\phi)} \mu(X^{-1}(C)) \max_{a\in A} \phi(C)_a.$$

Lower bound. Suppose $f = (X, Y) \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$. For every $\mu \in \mathcal{M}$ and every active cell C, we have

$$\mathbb{E}_{\omega\sim\mu}[\pi^{\phi}(X(\omega)) \cdot Y(\omega) \mid \omega \in X^{-1}(C)] \geq \mathbb{E}_{\omega\sim\mu}[\pi^{\phi}(X(\omega)) \cdot \phi(C) \mid \omega \in X^{-1}(C)] = \max_{a\in A} \phi(C)_a.$$

Since $\mu(X^{-1}(C)) = 0$ for each $\mu \in \mathcal{M}$ and $C \notin \mathcal{A}(\phi)$, we have

$$U_{\mathcal{M}}(f, \pi^{\phi}) \geq \inf_{\mu\in\mathcal{M}} \sum_{C\in\mathcal{A}(\phi)} \mu(X^{-1}(C)) \max_{a\in A} \phi(C)_a.$$

Taking the infimum over $f \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$ gives $\inf_{f\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f, \alpha^{\mathcal{C}}(\phi)) \geq W(\phi)$.

Upper bound for alternative policies. Now fix any targeted policy π. For each $f = (X, Y) \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$, define a bounded Borel measurable outcome process $\tilde{Y}$ by setting $\tilde{Y}(\omega) = \phi(C)$ whenever $\omega \in X^{-1}(C)$ for some $C \in \mathcal{A}(\phi)$, and setting $\tilde{Y}(\omega) = 0$ otherwise; let $\tilde{f} = (X, \tilde{Y})$. Since $\mu(X^{-1}(C)) = 0$ for each $\mu \in \mathcal{M}$ and $C \notin \mathcal{A}(\phi)$, we have $\ell^{\mathcal{M}}_{a,C}(f) = \ell^{\mathcal{M}}_{a,C}(\tilde{f})$ for all $a \in A$ and $C \in \mathcal{C}$; then $\underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(\tilde{f}) = \phi$. Then for every $\mu \in \mathcal{M}$,

$$\mathbb{E}_{\omega\sim\mu}[\pi(X(\omega)) \cdot \tilde{Y}(\omega)] \leq \sum_{C\in\mathcal{A}(\phi)} \mu(X^{-1}(C)) \max_{a\in A} \phi(C)_a;$$

$$\Rightarrow \inf_{\hat{f}\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(\hat{f}, \pi) \leq U_{\mathcal{M}}(\tilde{f}, \pi) \leq \inf_{\mu\in\mathcal{M}} \sum_{C\in\mathcal{A}(\phi)} \mu(X^{-1}(C)) \max_{a\in A} \phi(C)_a.$$

Taking the infimum over $f = (X, Y) \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$ gives

$$\inf_{\hat{f}\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(\hat{f}, \pi) \leq W(\phi).$$

Combining the two bounds proves

$$\inf_{f\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f, \alpha^{\mathcal{C}}(\phi)) \geq \inf_{f\in\underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f, \pi)$$

for every targeted policy π and every explanation $\phi \in \underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(F)$.

Strictness. Fix a targeted policy π. Choose $x^* \in \mathcal{X}$ and an action a such that $\pi_a(x^*) < 1$. Fix any prior $\mu \in \mathcal{M}$ and a support point ω_0 of μ. Choose a nondegenerate closed interval $J \subseteq [0,1]$ such that $[0,1] \setminus J$ contains a neighborhood of ω_0. By Lemma 3, there exists a continuous surjection $q : [0,1] \to \mathcal{X}$ with $q(0) = q(1) = x^*$. Define a continuous onto covariate process $\hat{X}$ by setting $\hat{X}(\omega) = x^*$ for $\omega \notin J$ and $\hat{X}(\omega) = q((\omega - a_J)/(b_J - a_J))$ for $\omega \in J$, where $J = [a_J, b_J]$. Let $\hat{Y}(\omega) = e_a$ for each $\omega \in [0,1]$, and denote $\phi = \underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}(\hat{X}, \hat{Y})$ and $\pi^\phi = \alpha^{\mathcal{C}}(\phi)$. For each $C \in \mathcal{A}(\phi)$, $\phi(C)_a = 1$ and $\phi(C)_b = 0$ for every $b \neq a$, so $\pi^\phi(x) = e_a$ for each $x \in C$. Then for each $f = (X, Y) \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)$,

$$U_{\mathcal{M}}(f, \alpha^{\mathcal{C}}(\phi)) = \sum_{C \in \mathcal{A}(\phi)} \mu(X^{-1}(C)) \mathbb{E}_{\omega \sim \mu}[e_a \cdot Y(\omega) \mid \omega \in X^{-1}(C)] \geq \sum_{C \in \mathcal{A}(\phi)} \mu(X^{-1}(C)) = 1.$$

But for $\hat{f} = (\hat{X}, \hat{Y})$,

$$U_{\mathcal{M}}((\hat{X}, \hat{Y}), \pi) \leq \mathbb{E}_{\omega \sim \mu}[\pi_a(\hat{X}(\omega))] \leq \mu([0,1] \setminus J)\pi_a(x^*) + \mu(J) < 1,$$

because $\mu([0,1] \setminus J) > 0$ and $\pi_a(x^*) < 1$. Hence

$$\inf_{f \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f, \alpha^{\mathcal{C}}(\phi)) > \inf_{f \in \underset{\sim}{\Gamma}^{-1}_{\mathcal{C},\mathcal{M}}(\phi)} U_{\mathcal{M}}(f, \pi).$$

Thus $\underset{\sim}{\Gamma}_{\mathcal{C},\mathcal{M}}$ is robustly useful. ■